APPROXIMATE LIKELIHOOD FOR DEPENDENT NETWORKS AND HYPERLINK PREDICTIONS

BY

YUBAI YUAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Doctoral Committee:

Professor Annie Qu, Chair
Professor Xiaofeng Shao
Associate Professor Xiaohui Chen
Assistant Professor Yun Yang

Abstract

Network data has arisen as one of the most common forms of information collection. This is due to the fact that the scope of studies not only focuses on subjects alone, but also on the relationships among subjects. In this thesis, we address two major challenges in the network analysis.

In the first part of the thesis, we focus on the detection of community structure in the network. In practical, within-community members are more likely to be connected than between-community members, which is also reflected in that the edges within a community are intercorrelated. However, existing probabilistic models for community detection such as the stochastic block model (SBM) are not designed to capture the dependence among edges. In the first part, we propose a novel community detection approach to incorporate intra-community dependence of connectivities through the Bahadur representation. The proposed method does not require specifying the likelihood function, which could be intractable for correlated binary connectivities. In addition, the proposed method allows for heterogeneity among edges among different communities. In theory, we show that incorporating correlation information can achieve a faster convergence rate compared to the independent SBM, and the proposed algorithm has a lower estimation bias and accelerated convergence speed compared to the variational EM. Our simulation studies show that the proposed algorithm outperforms the existing variational EM algorithm assuming conditional independence among edges. We also demonstrate the application of the proposed method to agricultural product trading networks from different countries.

In the second part, we focus on the joint prediction of pairwise link and hyperlink under multilayer networks to incorporate high-order relations in network, which are not considered in the traditional graph representation models which only predict two-way pairwise relations. We propose a novel joint network embedding approach on simultaneously encoding pairwise links and hyperlinks onto a latent space to capture the dependency between pairwise and multi-way links, which

allows inference of potential unobserved hyperlinks. The major advantage of the proposed embedding procedure is that it incorporates both the pairwise relationships and subgroup-wise structure among nodes to utilize high-order network information. In addition, the proposed method introduces the hierarchical dependency among links to infer potential hyperlinks, and leads to a better link prediction. In theory, we establish the estimation consistency for the proposed embedding approach, and provide a faster converge rate compared to hyperlink prediction using pairwise links only. Numerical studies on both simulation settings and Facebook ego-network show that the proposed method improves both hyperlink and pairwise link predictions accuracy compared to the existing link prediction methods.

Dedicated to my parents, for their unconditional love and support.

Acknowledgments

First and foremost I would like to express my sincerest gratitude to my advisor Professor Annie Qu for her continuous support, encouragement and inspiration. Her great creativity, critical thinking and knowledge show me the way to great research in statistics; her consistent patience, passion and courage motivate me to overcome difficulties, build self-discipline and live each day earnestly. I would not start my Ph.D. study without her great support and suggestion. I cannot be more appreciative of her heuristic and persistent guidance which builds my research topics, broadens my horizon in statistics and machine learning, and drastically improves my academic skills. She persistently warms my heart via countless helps for both of my life and career development during the whole three years. I cannot imagine my doctoral research and the completion of this thesis without her mentorship.

Special thanks to my committee members Professor Xiaofeng Shao, Professor Xiaohui Chen, and Professor Yun Yang for their support in my career development and insightful comments and suggestions to my thesis research. Furthermore, I own my sincere thanks to Dr. Haoda Fu, Yujia Deng, Qi Xu, Diqing Li, Yanqing Zhang, Professor Xuan Bi and Professor Xiwei Tang for being wonderful collaborators. I have learned a lot from their deep insight and great knowledge.

My thanks are given to the faculty and staff members in the Statistics Department for shaping much of my knowledge and understanding about statistics with their excellent courses and providing a wonderful environment for my doctoral life. I am also grateful to all my fellows and lab members, especially Fei Xue, Yujia Deng, Yanqing Zhang, Diqing Li, Qi Xu, Jiuchen Zhang, Yuexia Zhang. It is my great fortune to have their company filling with warmth and joy. Thanks to my fellow Wenzhuo Zhou for his presence in my life and memorable friendship.

Finally, I would like to thank my family for their constant love, encouragement, support, and most importantly, for being in my life.

Contents

Chapter 1

Introduction

Network data has arisen as one of the most common forms of information collection, due to the fact that the scope of studies not only focuses on subjects alone, but also on the complex relations or associations among interacting units in a system. Networks consist of two components: (1) nodes or vertices corresponding to basic units of a system, and (2) edges representing connections between nodes. These two main components can have various interpretations under different contexts of applications. For example, nodes might be humans in social networks; molecules, genes, or neurons in biology networks; or web pages in information networks. Edges could be friendships, alliances, URLs, or citations. In this thesis, we propose novel methods incorporating more structured information in network data. We mainly focus on two research areas: community detection and hyperlink prediction.

1.1 Community Detection

For network data analyses, identifying communities is essential to provide deep understanding of relationships among nodes within a community and between communities to address scientific, social and political problems [115, 18, 44, 114, 86, 66, 78]. In terms of other applications, community detection plays an important role in decomposing original large-scale network structures [119, 111, 93] into several subnetworks with more simplified structures [27], and facilitates scalable computation for further analyses.

Under the statistics framework, the popular stochastic block model (SBM) [51] assumes the observed network with community structure is generated from an underlying generating process and the node clustering is achieved through the maximum likelihood method. The core assumption for the SBM and its variants is that connectivities are conditional independent given their communi-

ties. However, the conditional independency assumption typically does not hold in practice and the high-order clustering information based on the dependency among connectivities might be lost. In Chapter 2, we propose a novel community detection method to detect the joint community structures among multiple networks. The proposed method can simultaneously integrate the marginal and correlation information from edge connectivities to distinguish communities from each other by utilizing a truncated Bahadur representation [14].

1.2 Hyperlink Prediction

In many applications of network data, the complex interactions among multiple nodes are often presented in multiway relations and can be expressed by a hypergraph [26, 34, 92]. Consequently, hyperlink prediction has become one important aspect in network analysis, which infers potential high-order associations of a hypergraph. Statistically, hyperlink prediction further expands link prediction from two-way relations to multiway relations. In contrast to pairwise links reflecting two-way concordance, hyperlinks focus on a joint concordance among a group of nodes with many potential subgraph configurations, which could be hidden or unobserved in practice. In addition, the configuration space for hyperlink prediction becomes much larger than that for pairwise link prediction, attributed to the nature of hyperlinks with combinations of multiple nodes. All of these make hyperlink prediction very challenging, which also motivate us to develop innovative models to improve link prediction through integrating high-order structures of networks.

In Chapter 3, we develop a novel approach to encode the potential subgroup structure onto a latent space, capturing the multiway link dependency to infer potential unobserved hyperlinks. The major advantage is that hyperlink prediction can be performed for high-order interactions through observed pairwise links in addition to the unobserved high-order subgroup structure, where the subgroup structure enhances hyperlink prediction by borrowing information from the within-subgroup dependency. A major novelty of the proposed research lies in the joint modeling of observed two-way and hidden multiway relations of network analysis.

Chapter 2

Community Detection with Dependent Network Connectivities

2.1 Introduction

There are increasing researches on scientific complex systems involve multiple networks [60, 116], where each individual network exhibits heterogeneous features through edge weights or edge density. However, these edges are also interconnected by underlying similarities such as shared network structures.

In this chapter, our goals are to detect the common community structures among multiple networks, which are motivated by sociology or neuroscience applications. One particular application for this type of data structure is from neuroimaging, where neuron connectivities in the brain are presented as network for each individual. Although the network connectivities vary for different subjects, it is also of scientific interest to identify the network community structure of brain's anatomical regions shared by all subjects, which is associated with functionally-specialized areas or general cognitive functions [70, 67, 20, 8, 77]. In addition, similar type of data structure can be found in the literature of international trading data which consists of a number of single trading network among countries, where each trading network corresponds to a specific product [38, 110]. Identifying the underlying trading groups of nations governed by their geographical and socio-economical similarity [110, 17] can be of political or business interest.

The network community detection can be summarized in the following two main categories. The first approach is the spectral method [102, 40, 12], which recovers dense connectivities through the low-rank approximation of the adjacent matrix of the network. The spectral methods on node clustering can be extended from a single network to multiple networks setting [109, 113, 21, 70]. One critical drawback of the spectral method is that it lacks stability to achieve lowest misclassification error, especially when the networks are sparse or the degrees of nodes are high (e.g., hubs [62]).

However, the estimation from spectral methods can serve as good initial values for other network analyses.

The second approach is to search a partition of nodes which optimizes a global criterion over all possible partitions, where the criterion function measures the goodness of fit of a partition such as the modularity [83] or likelihood function under a certain statistical model for observed networks. One particular network model is the stochastic block model (SBM) [51]. Alternatively, the likelihood-based community detection methods are proposed such as profile likelihood [22, 106, 68, 10], degree-corrected blockmodel [56, 122] incorporating the heterogeneity of nodes' degrees, latent position model [48, 49] considering latent distance to handle overlapping communities [4, 16]. Under the multiple networks setting, [67] develops the likelihood approach to identify the shared communities. In general, optimizing the likelihood criterion could be computationally intensive, and may suffers from ignoring small communities [43].

The common key assumption for the above methods is that connectivities are conditional independent given the membership of nodes. However, the network data are likely dependent among connectivities, which are also considered in several random network modelings [50, 65, 58, 31]. For community detection, the conditional independency assumption typically does not hold in practice and therefore could lead to a misspecified model [97, 11, 112]. For example, friendships within a social community or functional connectivities in brain networks tend to be highly correlated.

In addition, under conditional independence, the community structure can only be identified based on the marginal mean discrepancy of connectivities between within-communities and across-communities. Specifically, as a fundamental assumption of the independent SBM, the marginal mean discrepancy is required to be greater than a sharp threshold to guarantee community detectability ([75, 79]). However, the marginal mean discrepancy assumption might not hold, while the correlations among edges could be non-negligible and highly informative in identifying community structures. We show that the proposed method is able to incorporate the correlation information to achieve consistent community detection when the marginal mean discrepancy is insignificant.

More recently, the SBM has been extended to address the heterogeneity feature of within-community for multiple network samples. For example, [108, 90] apply a fixed-effect model

through an independent intercept without incorporating information from other networks. Alternatively, a random-effects model is proposed to incorporate heterogeneity [91, 121], which borrows information from multiple networks. However, both of these approaches require the specification of a distribution for the random effects. In addition, an EM-type algorithm is implemented to integrate out the random-effects, [91, 121] which could be computationally expensive when the size of the community or the network size is large.

In this chapter, we propose a novel community detection method to jointly model community structures among multiple networks. The proposed method can simultaneously incorporate the marginal and correlation information to differentiate within-community and between-community connectivities. The key idea is to approximate the joint distribution of correlated within-community connectivities by using a truncated Bahadur representation [14]. Although the approximate likelihood function is not the true likelihood, it is able to maximize the true community memberships and serves as a tighter lower bound to the true likelihood compared with the independent SBM likelihood. Consequently, we identify communities via maximizing the approximate likelihood function, which also serves as a discriminative function for membership assignments of nodes. In particular, within-community correlations provide an additional community-concordance measurement to capture high-order discrepancy between within-community and across-community networks, and therefore increase discriminative power to identify communities.

The main advantages and contributions of the proposed method can be summarized as follows. The proposed method incorporates correlation information among connectivities to achieve more accurate community detection than the variational EM method using marginal information only. The improvement of the proposed community detection method is especially powerful when the marginal information is relatively weak in practice. In addition, compared to the existing random-effects model, the proposed method is more flexible in modeling the heterogeneity of communities for multiple networks and heterogeneity of correlations among edges. Furthermore, it does not require a distribution specification among within-community connectivities.

In addition, we establish the consistency of the community estimation for the proposed approximate likelihood under a general within-community edge correlation structure and show that the proposed method achieves a faster convergence rate of membership estimation compared to the

independent likelihood. In terms of computational convergence, the proposed algorithm achieves a lower estimation bias and a faster convergence rate compared to the variational EM algorithm at each iteration via incorporating additional correlation information. The theoretical development in this paper is nontrivial, since establishing membership estimation consistency is more challenging under the framework of conditional dependency among edges compared to the existing ones assuming the conditional independent model. Furthermore, we show that the convergence of the variational EM algorithm [74] is a special case of our method under the conditional independent SBM.

Computationally, we develop a two-step iterative algorithm which is not sensitive to initial values as in the standard variational EM algorithm. In addition, compared to the existing fixed-effects SBM with independent intercepts or the random-effects SBM, the proposed method has lower computational complexity, as it does not involve integration of random effects as in [91], or estimating the fixed effects for each network as in [90]. Simulation studies and a real data application also confirm that the proposed method outperforms the existing variational EM significantly, especially when the marginal information of observed networks is weak. This chapter is organized as follows: Section 2.2 introduces the background of the proposed method. Section 2.3 introduces the proposed method to incorporate correlation information for community detection. Section 2.4 provides an algorithm and implementation strategies. Section 2.5 illustrates the theoretical properties of the proposed method. Section 2.6 demonstrates simulation studies, and Section 2.7 illustrates an application to world agricultural products trading data. The last section provides conclusions and some further discussion.

2.2 Background and Notation

In this section, we provide background and notation of the proposed community detection. The stochastic block model (SBM) [51] is a form of hierarchical modeling which captures the community structure for networks. Consider M symmetric and unweighted sample networks $\boldsymbol{Y} = \{\boldsymbol{Y}^m\}_{m=1}^M = \{(Y_{ij}^m)_{N\times N}\}_{m=1}^M$ with N nodes for K communities. Let $\{z_i\}_{i=1}^N$ be the membership for each node and $z_i \in \{1, 2, \cdots, K\}$, and denote the membership assignment matrix $\boldsymbol{Z} = \{(Z_{iq})_{n\times K}\} \in \{0,1\}^{N\times K}$, where $Z_{iq} = \mathbb{1}\{z_i = q\}$. Here $\boldsymbol{Z}$ has exactly one 1 in each

row and at least one 1 in each column for no-null communities. The unknown membership $z_i \in \{1, 2, \cdots, K\}$ can be modeled as a latent variable from a multinomial distribution:

$$z_i \sim Multinominal(1, \alpha_i),$$

where $i = 1, \cdots, N$, $\alpha_i = \{\alpha_{i1}, \cdots, \alpha_{iK}\}$ and $\sum_{k=1}^{K} \alpha_{ik} = 1$. Given the membership of nodes, the observed edges between two nodes $\{(Y_{ij}^m)_{n\times n}\}_{m=1}^{M}$ typically follow a Bernoulli distribution:

$$f_{ql}(Y_{ij}^m) := P(Y_{ij}^m | z_i = q, z_j = l) \sim Bern(\mu_{ql}), \text{ for } i, j \in \{1, \cdots, N\},\ q, l = 1, \cdots, K, \quad (2.1)$$

where μ_{ql} is the probability of nodes i and j being connected.

For the heterogeneous stochastic blocks model, the marginal mean μ_{ql} for each block in the mth network can be modeled as a logistic model to incorporate heterogeneity among edges:

$$\mu_{ql}^m = exp(\beta_{ql} x_{ij}) / \big\{1 + exp(\beta_{ql} x_{ij})\big\}, \quad (2.2)$$

where $(x_{ij})_{N\times N}$ are edge-wise covariates, and edges within the same community preserve homogeneity by sharing a block-wise parameter β_{ql}. The joint likelihood function can be decomposed into a summation of edge-wise terms following the conditional independence assumption:

$$logP(\boldsymbol{Y}; \boldsymbol{Z}) = \sum_{m=1}^{M} \sum_{q=1}^{K} \sum_{i=1}^{N} Z_{iq} log\alpha_q + \sum_{m=1}^{M} \sum_{q,l=1}^{K} \sum_{i<j}^{N} Z_{iq} Z_{jl} f_{ql}(Y_{ij}^m; \beta_{ql}). \quad (2.3)$$

The latent membership $\boldsymbol{Z}$ is estimated by $E(\boldsymbol{Z}|\boldsymbol{Y})$ through the maximum likelihood estimator of model parameters $\Theta = \{\beta_{ql}; q, l = 1, \cdots, K;\ \alpha_q; q = 1, \cdots, K\}$ in (2.3). However, the classical EM algorithm is not applicable here, because the conditional distribution $P(\boldsymbol{Z}|\boldsymbol{Y}) = \dfrac{P(\boldsymbol{Y}; \boldsymbol{Z})}{\sum_{\boldsymbol{Z}} P(\boldsymbol{Y}; \boldsymbol{Z})}$ becomes intractable in the expectation step.

The variational EM algorithm [74, 53] is one of the most popular inference methods, and can be applied to approximate the likelihood $P(\boldsymbol{Z}|\boldsymbol{Y})$ by a complete factorized distribution $R(\boldsymbol{Z}, \boldsymbol{\tau}) = \prod_{i=1}^{N} h(Z_i; \tau_i)$, where $h(\cdot)$ denotes a multinomial distribution, $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_N)$ and $\tau_i = (\tau_{i1}, \cdots, \tau_{iK})$ is a probability vector such that $\sum_{q=1}^{K} \tau_{iq} = 1$. In the expectation step, the likelihood $logP(\boldsymbol{Y}; \boldsymbol{Z})$ is

averaged over $R(\boldsymbol{Z})$ such that for any $\boldsymbol{\tau}$, $E_{R(\boldsymbol{Z},\boldsymbol{\tau})}\big\{logP(\boldsymbol{Y};\boldsymbol{Z})\big\} \leq E_{P(\boldsymbol{Z}|\boldsymbol{Y})}\big\{logP(\boldsymbol{Y};\boldsymbol{Z})\big\}$ where,

$$E_{R(\boldsymbol{Z},\boldsymbol{\tau})}\big\{logP(\boldsymbol{Y};\boldsymbol{Z})\big\} = -\sum_{m=1}^{M}\sum_{q=1}^{K}\sum_{i=1}^{N}\tau_{iq}log\tau_{iq} + \sum_{m=1}^{M}\sum_{q=1}^{K}\sum_{i=1}^{N}\tau_{iq}log\alpha_q + \sum_{m=1}^{M}\sum_{q,l=1}^{K}\sum_{i<j}^{N}\tau_{iq}\tau_{jl}f_{ql}(Y_{ij}^m).$$

Instead of directly maximizing $E_{P(\boldsymbol{Z}|\boldsymbol{Y})}\big\{logP(\boldsymbol{Y};\boldsymbol{Z})\big\}$, the variational EM approach alternatively maximizes its lower bound $E_{R(\boldsymbol{Z},\boldsymbol{\tau})}\big\{logP(\boldsymbol{Y};\boldsymbol{Z})\big\}$ over model parameters Θ and variational parameters $\boldsymbol{\tau}$, and clusters nodes by $\boldsymbol{\tau}$ through $\hat{z}_i = \text{argmax}_k\{\hat{\tau}_{ik}, k = 1, \cdots, K\}$.

Throughout this paper, we consider the conditional version of SBM (CSBM) [22, 102, 33], where the true membership $\boldsymbol{Z}^*$ is fixed. The conditional stochastic block model framework assumes conditional independence among edges, i.e., $Y_{i_1j_1}^m$ and $Y_{i_2j_2}^m$ are independent given nodes' membership $z_{i_1}, z_{i_2}, z_{j_1}, z_{j_2}$, and the corresponding log-likelihood of observed sample networks is:

$$logL_{ind}(\boldsymbol{Y}|\boldsymbol{Z}) = \frac{1}{M}\sum_{m=1}^{M}\sum_{q,l=1}^{K}\sum_{i<j}^{N}Z_{iq}Z_{jl}\Big\{y_{ij}^m log\ \mu_{ql} + (1-y_{ij}^m)log\ (1-\mu_{ql})\Big\}. \tag{2.4}$$

The above log-likelihood can serve as a discriminant function in clustering membership $\boldsymbol{Z}$ in that if $logL_{ind}(\boldsymbol{Y}|\boldsymbol{Z_1}) > logL_{ind}(\boldsymbol{Y}|\boldsymbol{Z_2})$ given two membership assignments $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$, then $\boldsymbol{Z_1}$ is preferred over $\boldsymbol{Z_2}$, since the likelihood for the observed sample networks is higher. Naturally, $\boldsymbol{Z}^*$ can be estimated by

$$\hat{\boldsymbol{Z}} = \underset{\boldsymbol{Z}}{\text{argmax}}\ logP_{ind}(\boldsymbol{Y}|\boldsymbol{Z}).$$

The SBM in (2.4) allows one to differentiate within-community and between-community nodes via utilizing only the marginal information, in that the average connectivity rates within-communities are higher than those between-communities. However, the underlying conditional independence assumption among edges is too restrictive and practically infeasible. In most community detection problems it is common that edges within communities are more correlated. For example, social connections among friends are highly correlated in social networks. However, the dependency among edges is not captured by the traditional SBM, which could lead to significant information loss of the community structure.

2.3 Methodology

2.3.1 Community Detection with Dependent Connectivity

In this chapter, we incorporate within-community correlation to improve accuracy and efficiency in identifying communities, in addition to utilizing the edges' marginal mean information, since within-community dependency contains additional information regarding the membership of nodes. This is especially effective when the marginal mean is not informative in differentiating between and within communities' connectivity.

In this section, we propose an approximate likelihood function to capture the dependency among within-community edges. We assume that each observed sample networks $Y^m_{n\times n}$ follow a multivariate binary distribution $P(Y^m)$ define in (2.1), where there exists correlations among within-community edges. Specifically, the correlation among a pair of edges $(Y^m_{i_1j_1}, Y^m_{i_2j_2})$ within a community is denoted as: $corr(Y^m_{i_1j_1}, Y^m_{i_2j_2}) = \rho_q(i_1, i_2, j_1, j_2) \in (-1, 1)$ given nodes z_{i_1}, z_{i_2}, z_{j_1} and z_{j_2} are in the same community q, where $1 \le i_1 < j_1 \le N, 1 \le i_2 < j_2 \le N, (i_1, j_1) \ne (i_2, j_2)$ and $q = 1, \cdots, K$. Note that correlations among each pair of edges are allowed to be different. Equivalently, the edges in community k show strong dependency only when

$$\sum_{i<j;u<v}^{N} Z_{ik}Z_{jk}Z_{uk}Z_{vk}\rho_q(i, j, u, v)\hat{y}^m_{ij}\hat{y}^m_{uv}$$

is large, where $\hat{y}^m_{ij}$ and $\hat{y}^m_{uv}$ are standardization of $Y^m_{i_1j_1}$ and $Y^m_{i_2j_2}$ by adjusting their marginal means. Note that correlations among edges can be incorporated for community detection on multiple networks. For example, [90, 91] utilize random effects to model the heterogeneity of the connectivities for an individual network, which leads to a positive correlation among the edges within the same community. In practice, both positive and negative correlations among edges could occur. For example, the positive pairwise correlations among edges are more likely to produce star or triad relations, and are widely observed in social networks [101, 100]. The negative correlation among edges could occur when there are the competitive relations among local retailers within the same geographical region.

The exponential random graph models (ERGMs) [100] can incorporate dependency among

edges. However, it differs from the block-dependency modeling in terms of capturing different network features. Specifically, the ERGMs characterize specific interest subgraphs in the network through the edge dependency. In addition, nodes are equivalent or exchangeable under the ERGMs such that realizations of subgraphs are assumed to be independent and serve as individual samples for the model. In contrast, the block-dependency approaches associate the edge dependency with the underlying community structure to capture the overall correlation intensity within communities instead of capturing the dependent structure in specific subgroups. In addition, nodes in the block-dependency model are not exchangeable as they might belong to different communities.

2.3.2 Approximate Likelihood

In this section, we propose an informative approximation of the true log-likelihood to cluster $\boldsymbol{Z}$ via incorporating interactions among edges within a community in addition to marginal mean information. This is because the exact joint likelihood function of correlated binary distribution $P(Y^m)$ is computationally intractable. Specifically, we construct an approximate likelihood as a substitute of the true likelihood by facilitating the Bahadur representation [14]. That is, we retain the low-order dependency information among edges within-communities and discard the high-order dependency for computational efficiency. Although the approximate likelihood is not a true likelihood, it still serves the purpose of estimating the membership of nodes.

Consider T dependent binary random variables, then the joint likelihood can be represented through the Bahadur representation:

$$P(Y_1 = y_1, \cdots, Y_T = y_T) = \prod_{j=1}^{T} \mu_j^{y_j}(1-\mu_j)^{1-y_j}\big[1 + \sum_{1\le j_1<j_2\le T} \rho_{j_1 j_2}\hat{y}_{j_1}\hat{y}_{j_2} + \sum_{1\le j_1<j_2<j_3\le T} \rho_{j_1 j_2 j_3}\hat{y}_{j_1}\hat{y}_{j_2}\hat{y}_{j_3} + \cdots + \rho_{12\cdots T}\hat{y}_1\hat{y}_2\cdot\cdot\hat{y}_T\big], \tag{2.5}$$

where

$$\mu_j = E(Y_j),\ \hat{y}_j = \frac{y_j - E(y_j)}{\sqrt{E(y_j)(1-E(y_j))}}, \tag{2.6}$$

and

$$\rho_{j_1 j_2} = E(\hat{y}_{j_1}\hat{y}_{j_2}),\ \rho_{j_1 j_2 j_3} = E(\hat{y}_{j_1}\hat{y}_{j_2}\hat{y}_{j_3}), \cdots, \rho_{12\cdots T} = E(\hat{y}_1\hat{y}_2 \cdot \cdot \hat{y}_T).$$

The idea of Bahadur representation is to approximate the joint distribution of dependent binary random variables as a function of moments with a sequential order. For the community detection problem, the binary random variables represent within-community edges, and the corresponding joint distribution can be explicitly decomposed into a marginal part and a correlation part. The marginal part consists of all the marginal mean μ_{ij} for each edge, which can be directly modeled through the dependency of the mean on covariates as in (2.2). The correlation part consists of interactions among all possible pairwise-associations of normalized edges, which add correlation information beyond a conditional independence likelihood model. Note that the conditional independence model is a special case of the proposed model when the correlation is zero, and the corresponding Bahadur representation collapses to a marginal part only, which is equivalent to the $\log L_{ind}(\boldsymbol{Y}|\boldsymbol{Z})$ in (2.4).

There are two major challenges in applying the Bahadur representation to model the interactions among within-community edges. First, the dimension of correlation parameters could be high if all the high-order interactions in (2.5) are incorporated, and this could lead to an increasing computational demand as the size of community grows. To solve this problem, we retain all the second-order interactions, but ignore interactions for higher orders beyond the second order, since the pairwise interactions among edges could be most important. In addition, we can further reduce the number of parameters via a homogeneous correlation structure such that all the pairwise correlations in each community are assumed to be the average within-community correlation given the sign of correlations are consistent in a community, which can be simplified as an exchangeable correlation structure. The rationales of this simplification are based on the following. First, the pairwise correlation parameter $\rho_q(i_1, i_2, j_1, j_2)$ is a nuisance correlation parameter to enhance clustering. Second, both the numerical experiments and theoretical findings show that the density of pairwise correlation among within-community edges plays a more important role than the intensity of the correlation in affecting clustering performance.

The second challenge is that the range of the correlation coefficient could be constrained by the marginal means [39]. Consequently, the correlation parameter space is more restrictive if the vari-

ability of marginal means among edges is large. Nevertheless, our primary goal is to construct an objective function which can incorporate information from the marginal mean and correlations of edges within-community, and the objective function is not necessarily the true likelihood function. In the proposed method, we instead construct an approximate likelihood which is more flexible for incorporating highly dependent communities while still achieving computational efficiency.

Specifically, we construct an approximate likelihood $\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z})$ incorporating correlated within-community edges as follows:

$$\begin{aligned}\log\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z}) =&\frac{1}{M}\Bigg\{\sum_{m=1}^{M}\sum_{q,l=1}^{K}\sum_{i<j}^{N} Z_{iq}Z_{jl}\Big\{y_{ij}^{m}\log\,\mu_{ql}+(1-y_{ij}^{m})\log\,(1-\mu_{ql})\Big\}\\ &+\sum_{m=1}^{M}\log\Big\{1+\sum_{k=1}^{K}\frac{1}{2}\max\{\sum_{\substack{i<j;u<v\\(i,j)\neq(u,v)}}^{N} Z_{ik}Z_{jk}Z_{uk}Z_{vk}\rho_{ijuv}\hat{y}_{ij}^{m}\hat{y}_{uv}^{m},0\}\Big\}\Bigg\},\end{aligned} \tag{2.7}$$

where μ_{ql} and $\hat{y}_{ij}^{m}$ are formulated in (2.2) and (2.6), and ρ_{ijuk} is the pairwise correlation between $\hat{y}_{ij}^{m}$ and $\hat{y}_{uv}^{m}$. Notice that the first term in (2.7) is the same as the marginal mean model, and the second term in (2.7) measures the concordance among edges within communities clustering $\boldsymbol{Z}$.

We denote the second term of (2.7) as

$$\log L_{cor}(\boldsymbol{Y}|\boldsymbol{Z}) = \frac{1}{M}\Bigg\{\sum_{m=1}^{M}\log\Big\{1+\sum_{k=1}^{K}\frac{1}{2}\max\{\sum_{\substack{i<j;u<v\\(i,j)\neq(u,v)}}^{N} Z_{ik}Z_{jk}Z_{uk}Z_{vk}\rho_{ijuv}\hat{y}_{ij}^{m}\hat{y}_{uv}^{m},0\}\Big\}\Bigg\}. \tag{2.8}$$

Compared with $\log L_{ind}(\boldsymbol{Y}|\boldsymbol{Z})$ in (2.4), the proposed $\log\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z})$ has more discriminative power over $\boldsymbol{Z}$, since it utilizes more information of the observed dependency within communities corresponding to clustering $\boldsymbol{Z}$. In addition, the nonnegativity of $\log L_{cor}(\boldsymbol{Y}|\boldsymbol{Z})$ ensure the fact that $\log\tilde{L}(\mathbf{Y}|\mathbf{Z}) \geq \log L_{ind}(\mathbf{Y}|\mathbf{Z})$ is guaranteed, which implies that adding additional correlation information among edges can be more informative given within-community correlation exists. This leads to higher classification accuracy and estimation efficiency through maximizing (2.8).

The key part of the proposed method is to predict memberships of nodes through the Bayes factor constructed by the proposed $\log\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z})$. Suppose the memberships of other nodes $\boldsymbol{Z}_{-i}$ are

known, then we classify node i based on the following Bayes factor:

$$\frac{\tilde{L}(\mathbf{Y}|\boldsymbol{Z}_{-\boldsymbol{i}}, Z_{iq}=1)}{\tilde{L}(\mathbf{Y}|\boldsymbol{Z}_{-\boldsymbol{i}}, Z_{ik}=1)} = exp\Big\{\log\tilde{L}(\mathbf{Y}|\mathbf{Z}_{-\boldsymbol{i}}, Z_{iq}=1) - \log\tilde{L}(\mathbf{Y}|\mathbf{Z}_{-\boldsymbol{i}}, Z_{ik}=1)\Big\}.$$

If the above Bayes factor > 1, then the probability of node i in community q is larger than that of community k. The Bayes factor can be further decomposed as:

$$\frac{\tilde{L}(\mathbf{Y}|\boldsymbol{Z}_{-\boldsymbol{i}}, Z_{iq}=1)}{\tilde{L}(\mathbf{Y}|\boldsymbol{Z}_{-\boldsymbol{i}}, Z_{ik}=1)} = \frac{L_{ind}(\mathbf{Y}|\boldsymbol{Z}_{-\boldsymbol{i}}, Z_{iq}=1)}{L_{ind}(\mathbf{Y}|\boldsymbol{Z}_{-\boldsymbol{i}}, Z_{ik}=1)} \frac{L_{cor}(\mathbf{Y}|\boldsymbol{Z}_{-\boldsymbol{i}}, Z_{iq}=1)}{L_{cor}(\mathbf{Y}|\boldsymbol{Z}_{-\boldsymbol{i}}, Z_{ik}=1)}, \tag{2.9}$$

which contains both the marginal ratio and the correlation ratio. It is clear that when the marginal information is weak in differentiating two communities, the marginal ratio is close to 1, and if the correlation ratio is informative, it can enhance the Bayes factor to improve community detection. In addition, the correlation ratio also serves as a correction to lower the estimation bias.

We illustrate the advantage of the proposed method in (2.8) over the conditional independent likelihood (2.4) using a simple numerical illustration. Specifically, we generate multiple networks based on the SBM with 30 nodes evenly split between two communities. The marginal means of within-community and between-community edges are the same at 0.5, implying that the marginal mean is not informative. We assume a true exchangeable correlation $\rho = 0.6$ for within-community edges. Figure 2.1 illustrates that the likelihood function changes as memberships of nodes change with some misclassified nodes. The left graph is based on the conditional independent SBM utilizing only marginal information, which does not differentiate the two communities at all due to weak marginal information. However, the proposed approximate likelihood in the right graph has high differentiation power for the nodes' memberships, and reaches maximum when the true memberships are selected.

2.4 Algorithm and Implementation

In this section, we propose a two-step algorithm to maximize the proposed approximate likelihood function. In addition, we provide implementation strategies to improve the stability and efficiency of the algorithm.

2.4.1 Algorithm

To estimate the true membership $\boldsymbol{Z}^*$ of nodes, we can ideally search through all the possible $\boldsymbol{Z}$ and choose the one with the largest $\log\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z})$. However, this becomes infeasible when the number of nodes N and the number of communities K increases. In the following, we propose an iterative two-step algorithm to maximize $\log\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z})$ in (2.7).

Algorithm 1

Step 1: Input an initial membership probability for each node: $\alpha_{iq}^{(0)},\ 1 \le i \le N,\ 1 \le q \le K$ through spectral clustering on individual sample networks.

Estimate each pairwise correlation $\{\rho(i,j,u,v)\}$ through the empirical estimator by $(\{Y_{ij}^m, Y_{uv}^m\}_{m=1}^M$.

Step 2: At the sth iteration, given $\{\beta_{ql}^{(s-1)},\ \rho_q^{(s-1)}\}_{q,l=1}^K$ and $\{\alpha_i^{(s-1)}\}_{i=1}^N$ from the $(s-1)$th iteration:

(i) Maximization: block-wise update $\beta_{ql}^{(s)}$ and $\rho_q^{(s-1)},\ q,l = 1,\cdots,K$;

(a) Obtain $\beta_{ql}^{(s)}$ through GEE with current membership as working correlation;

(ii) Expectation: given $\{\beta_{ql}^{(s)},\ \rho_q^{(s)}\}_{q,l=1}^K$, update $\{\alpha_i^{(s)}\}_{i=1}^N$:

$$\alpha_{iq}^{(s)} = \frac{\alpha_{iq}^{(s-1)}\tilde{L}(\mathbf{Y}|\alpha_{-i}^{(s-1)}, Z_{iq}=1)}{\sum_{k=1}^K \alpha_{ik}^{(s-1)}\tilde{L}(\mathbf{Y}|\alpha_{-i}^{(s-1)}, Z_{ik}=1)},\ i = 1,\cdots,N,\ q = 1,\cdots,K.$$

Step 3: Iterate until $\max_{1\le i\le N} |\alpha_i^{(s)} - \alpha_i^{(s-1)}| < \epsilon$.

Step 4: Obtain the membership z_i of clusters by

$$\{\alpha_i^{(s)}\}_{i=1}^N\text{: } z_i = \max_k\{\alpha_{i1}^{(s)},\cdots,\alpha_{iK}^{(s)}\},\ i = 1,\cdots,N.$$

Here we directly maximize the approximate likelihood instead of a true likelihood as in the EM algorithm. In the expectation step, we alternatively update membership of each node while fixing other nodes, where $\tilde{L}(\mathbf{Y}|\alpha_{-i}; Z_{ik})$ has the same formulation as $\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z})$ in (2.7) with $\{Z_{iq}\}_{N\times K}$ replaced by its expectation $\{\alpha_{iq}\}_{N\times K}$, except Z_{ik}. Note that α_{iq} is not the expectation under the true underlying joint distribution $P(Y,Z) = P(Y|Z)P(Z)$. Instead, it corresponds to the distribution defined by the approximate likelihood in (2.8). In the expectation step, the memberships are updated through the Bayes factor in (2.9) with the proposed $\tilde{L}(\mathbf{Y}|\mathbf{Z})$. In the maximization

step, we estimate the community-wise parameters β_{ql} through the generalized estimating equation where the working correlation is exchangeable structure given the current membership of nodes and estimated average correlation ρ_q. Note that the variational EM is a special case of the proposed algorithm if the correlation information is ignored and the conditional independent model in (2.4) is assumed.

2.4.2 Computation and Implementation:

To ensure computational stability, the community-wise parameters β_{ql} could be estimated through a simplified generalized estimation equation assuming an independent working correlation in algorithm 1. This is because the primary interest of community detection is classification accuracy, and the empirical studies show that correlation information plays a relatively minor role in parameter estimation.

We can achieve a better approximation to the true likelihood if higher-order moments are incorporated in the Bahadur representation in (2.6), which also increases its discrimination power. However, higher-order correlation could also increase the computational cost. Alternatively, we can recover partial higher-order interactions (e.g., the fourth order) derived from low order interactions (e.g., the second order). For example, consider four normalized edges $\hat{Y}^m_{i_1j_1}, \hat{Y}^m_{i_2j_2}, \hat{Y}^m_{i_3j_3}$ and $\hat{Y}^m_{i_4j_4}$ within the same community k with a positive fourth order correlation among them, we have

$$E\big(\hat{Y}^m_{i_1j_1}\hat{Y}^m_{i_2j_2}\hat{Y}^m_{i_3j_3}\hat{Y}^m_{i_4j_4}\big) \geq E\big(\hat{Y}^m_{i_1j_1}\hat{Y}^m_{i_2j_2}\big)E\big(\hat{Y}^m_{i_3j_3}\hat{Y}^m_{i_4j_4}\big) = \rho_{i_1j_1i_2j_2}\rho_{i_3j_3i_4j_4}. \tag{2.10}$$

To simplify notation, denote $(Z_{1k}Z_{2k}\hat{Y}^m_{12}, Z_{1k}Z_{3k}\hat{Y}^m_{13}, \cdots, Z_{2k}Z_{3k}\hat{Y}^m_{23}, \cdots, Z_{(N-1)k}Z_{Nk}\hat{Y}^m_{(N-1)N})$ as $(\gamma^m_1, \gamma^m_2, \cdots, \gamma^m_{N_0})$, where $N_0 = \frac{N^2-N}{2}$. Then the second-order interaction term for the community k in $L_{cor}(\boldsymbol{Y}|\boldsymbol{Z})$ is

$$\frac{\rho_k}{2}\sum_{\substack{i<j,u<v\\(i,j)\neq(u,v)}}^{N} Z_{ik}Z_{jk}Z_{uk}Z_{vk}\hat{y}^m_{ij}\hat{y}^m_{uv} = \rho_k\sum_{s<t}^{N_0}\gamma^m_s\gamma^m_t.$$

Based on (2.11) and given $\boldsymbol{Z}$, we can approximate the fourth-order interaction for community k

under the exchangeable correlation structure by its lower bound:

$$\sum_{\substack{s_1<t_1,s_2<t_2\\(s_1,t_1)\neq(s_2,t_2)}}^{N_0}\frac{E(\gamma_{s_1}^m\gamma_{t_1}^m\gamma_{s_2}^m\gamma_{t_2}^m)}{2}\gamma_{s_1}^m\gamma_{t_1}^m\gamma_{s_2}^m\gamma_{t_2}^m \geq \sum_{\substack{s_1<s_2,t_1<t_2\\(s_1,t_1)\neq(s_2,t_2)}}^{N_0}\frac{\rho_k^2}{2}\gamma_{s_1}^m\gamma_{t_1}^m\gamma_{s_2}^m\gamma_{t_2}^m = \Big(\rho_k\sum_{s<t}^{N_0}\gamma_s^m\gamma_t^m\Big)^2 - \rho_k^2\sum_{s<t}^{N_0}(\gamma_s^m\gamma_t^m)^2. \tag{2.11}$$

Note that the above lower bound of the fourth-order interaction can be calculated by the second-order interaction term in $L_{cor}(\boldsymbol{Y}|\boldsymbol{Z})$. Therefore, we can still incorporate higher-order terms in $log\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z})$ without additional computational cost. For other types of non-exchangeable correlation structures, we can incorporate partial higher-order correlation similarly as above. The main difference is that each pair of edges is associated with a specific correlation given a dependency structure. Therefore, the simplified lower bound for higher-order correlations such as (**??**) does not hold in general, and could have a more complex form depending on the specific correlation structure.

In the following, we also provide some guidelines for determining the number of communities K and initial membership of nodes. For a single network, the criterion-based methods choose K to maximize a certain probabilistic criterion such as the integrated likelihood [45, 37, 64], composite likelihood BIC [104] or modularity criterion [24]. In addition, spectral methods estimate K through the spectral property of the transformed adjacent matrix, such as a Laplacian matrix [82], non-backtracking matrix [25] or Bethe Hessian matrix [103]. In the hierarchical Bayesian framework, the number of communities is treated as a model parameter given a certain prior distribution and is jointly estimated with nodes' memberships using the MCMC [45, 84, 85]. For multiple networks, we can extend the above techniques to estimate a consensus number of communities combining observed realizations of the SBM from each individual network.

In the context of the proposed within-community dependent modeling, we can first perform the modularity-maximizing method or spectral clustering on each individual network to obtain K, then take the average of these individual estimated K, which can be treated as a consensus number of communities. The above procedure is sensible under two considerations. First, each sample network is a realization of the SBM so that the individual estimation of K is randomly distributed around the true underlying K. Thus the average of individual estimations provides an estimation

of K with low-bias and low-variance. Second, the spectral clustering or modularity methods are more favorable than other methods, due to their relatively low computational cost in estimating K. This is especially effective when the sample size of networks is large.

As an EM-type algorithm, the proposed optimization procedure can only guarantee the local maximum and requires multiple initializations to find the global maximum. In this paper, we obtain the membership initializations through spectral clustering on different sample networks, a benchmark algorithm for the traditional SBM. Spectral clustering is a model-free clustering algorithm and is able to provide a warm start for nodes' memberships.

2.5 Theoretical Results

In this section, we establish the consistency of the estimated nodes' membership based on the independent likelihood and the approximate likelihood approaches. In addition, we provide the computational convergence theorem for the proposed iterative algorithm in section 4. Compared to the independent likelihood approach, we show that the approximate likelihood approach leads to a computationally faster convergence rate regarding nodes' membership estimation.

2.5.1 Consistency of Nodes' Membership Estimation

In this subsection, we study the consistency of the maximization likelihood estimator for both the independent likelihood and the approximate likelihood at the population level. With the independence assumption among within-community edges, the consistency and convergence rate of the MLE estimator can be obtained by [29, 120]. However, the convergence property of the MLE remains unknown if there exists a local dependence among edges.

One significant distinction using the independence assumption if the edges are correlated is that the increasing number of nodes and number of edges do not necessarily guarantee a lower misclassification rate and computationally faster convergence. This is because the discrepancy between marginal means from within-community and between-community is not accumulated due to the pairwise correlation, though it can be accumulated through increasing the number of sample networks. However, we show that the proposed approximated approach is able to benefit from the

increasing number of nodes, and therefore achieves a faster computational convergence compared to the independent likelihood approach.

In the following theorems, we assume that edges within the same block have the same marginal mean such that $\mu_{z_i z_j} := E(Y_{ij}^m | i \in q, j \in l) = \eta_N c_{ij}$, where $\eta_N \in (0, 1]$ is a sparsity parameter controlling the average node degree. We denote that the true marginal means as $\Theta = \{\mu_{ql}, 1 \leq q < l \leq K\}$, and assume the following two regularity conditions regarding identifiability:

(C1). Suppose for every $q \neq q', 1 \leq q, q' \leq K$, there exists at least one $l \in \{1, \cdots, K\}$ such that $\mu_{ql} \neq \mu_{q'l}$. In addition, all the c_{ql} are bounded such that $c_{ql} \in [\zeta, 1 - \zeta], q, l = 1, \cdots, K$ with $\zeta > 0$.

For a more general case where the edgewise marginal means vary due to their varying covariates in (2.2), Theorem 2.1 and Theorem 2.2 can be generalized under the assumption similar to (C1) in that the edgewise marginal means in each community lie within K disjoint balls centered at vectors $(\mu_{1q}, \cdots, \mu_{Kq}),\ q = 1, \cdots, K$, respectively.

(C2). Community sizes from all sample networks are bounded above and below by $\kappa_1 N \leq |\{i \in \{1, 2, \cdots, N\} : Z_{iq}^* = 1\}| \leq \kappa_2 N,\ q = 1, \cdots, K$, where κ_1 and κ_2 are constants such that $0 < \kappa_1 < \kappa_2 < 1$.

In the following, we establish the consistency of membership estimation for both the independent likelihood approach and the proposed approximate likelihood approach. For the within-community edges, we define the edgewise second-order pairwise correlation density as

$$\lambda = \lambda_{ij}^m := \frac{\#|\{(u, v) : |cor(Y_{ij}^m, Y_{uv}^m)| > 0, Z_u = Z_v = k\}|}{N_k(N_k - 1)/2 - 1} \text{ for edge } Y_{ij}^m \text{ in community k}$$

where $k = 1, 2, \cdots, K$ and $N_k(N_k - 1)/2 - 1$ is the number of edges within community k for the sample network $\boldsymbol{Y}^m$. For simplicity, we assume the homogeneous second-order correlation density such that $\lambda_{ij}^m = \lambda$ for all the within-community edges. Here $\lambda \in [0, 1]$ severs as a counterpart of sparsity parameter η_N commensurate with edge correlation density, and determines the intensity of local dependency within a community. Specifically, $\lambda = 0$ indicates that within-community edges are all independent, while $\lambda = 1$ indicates that all edges within a community are pair-wisely correlated. In addition, correlation density λ is allowed to depend on the number of nodes, and increases such that it can model a more general class of correlation structure. For example, in a

hub structure, an edge is only correlated with those sharing the same hub nodes and the density $\lambda = \frac{N_k - 1}{N_k^2 - 1} = O_N(\frac{1}{N_k})$.

To establish asymptotic consistency for the proposed likelihood, we assume the sparsity of high-order correlation among within-community edges.

(C3). The number of third and fourth-order correlations defined in (2.5) among within-community edges do not exceed the order of the size of second-order correlations. Specifically, for edge Y_{ij}^m in community k, $\#|\{(i,j),(u_1,v_1),(u_2,v_2) : E(\hat{Y}_{ij}\hat{Y}_{u_1v_1}\hat{Y}_{u_2v_2}) \neq 0\}| \leq O_N(\lambda(N_k^2))$. In addition, $\#|\{(i,j),(u_1,v_1),(u_2,v_2),(u_3,v_3) : E(\hat{Y}_{ij}\hat{Y}_{u_1v_1}\hat{Y}_{u_2v_2}\hat{Y}_{u_3v_3}) \neq 0\}| \leq O_N(\lambda(N_k^2)), k = 1, 2, \cdots, K$.

In general, assume that the pairwise correlations among the within-community edges are sufficient to cover a broad class of Markov dependence modeling under the general exponential random graph model. This includes the most commonly used edge dependence configurations such as a star, a triangular shape subnetwork [81] and the k-triangles shape [89]. Although considering that the additional higher-order edge correlation improves the model's complexity, it could increase higher computational cost and instability. Empirically, it is sensible to assume that higher-order correlation only exists when second-order correlation already exists among edges, for the sake of identifiability and interpretability of the model. Otherwise, it could lead to the 'near degeneracy' [47] when a higher-order dependency masks a lower-order dependency.

Let $P_{Z^*} := \mathbb{P}(\cdot|Z = z^*; \Theta)$ denote the conditional distribution of edges given the true membership of nodes and true parameters.

Theorem 2.1. *Under the regularity conditions (C1)-(C3), we establish the convergence rate of the membership estimator z using the independent likelihood approach. That is, for every $t > 0$ and $z \neq z^*$,*

$$P_{Z*}\Big\{\frac{L_{ind}(\boldsymbol{Y}|\boldsymbol{Z} = z; \Theta)}{L_{ind}(\boldsymbol{Y}|\boldsymbol{Z} = z^*; \Theta)} > t\Big\} = \mathcal{O}\Big(\exp\Big\{-C_1 \frac{c^* r \eta_N N M}{1 + \rho\kappa_2\eta_N N \min(r, \kappa_2\lambda N)}\Big\}\Big), \tag{2.12}$$

where $r = \|z - z^*\|_0$ is the number of misclassified nodes up to the permutation labeling, ρ is the largest pairwise correlation among within-community edges, C_1 is a positive constant, and $c^* = \min_{(q,l),(q',l')} \{D_{KL}(c_{ql}||c_{q'l'}) : c_{ql} \neq c_{q'l'}\}$, where D_{KL} denotes the Kullback–Leibler divergence distance.

Given the convergence rate based on the independent likelihood ratio, we can characterize the convergence of its estimated node membership as following:

Corollary 2.1. *Under the same conditions given in Theorem 2.1, using the independent likelihood approach, for every* $t > 0$

$$P_{Z*}\Big\{ \sup_{\{z \neq z^*\}} \frac{L_{ind}(\boldsymbol{Y}|\boldsymbol{Z} = z; \Theta)}{L_{ind}(\boldsymbol{Y}|\boldsymbol{Z} = z^*; \Theta)} > t \Big\} = \mathcal{O}\Big(N \exp\Big\{ - \frac{c^* \eta_N NM}{1 + \eta_N \lambda N^2)} \Big\} \Big). \tag{2.13}$$

For the independent likelihood approach, the convergence rate depends on the number of sample network M, the marginal sparsity parameter η_N and the density of the pairwise correlation λ among within-community edges. If there is no pairwise correlation among edges, e.g., $\lambda = 0$, then the convergence rate in (2.13) increases to $\mathcal{O}_{N,M}\big\{N \exp(-\eta_N NM)\big\}$, which degenerates to the convergence rate established in [29] under the conditional independent modeling given constant marginal mean $\eta_N = 1$. In addition, the convergence of proposed node membership estimator can be guaranteed under the sparse growth rate of $q = (\log^c N)/N$ for some constant $c > 1$, which is consistent with the existing results in [1]. In general, The probability of true membership goes to 1 as M or the node size N increases under a relatively sparse pairwise correlation, e.g., $\eta_N \lambda N^2 = o_N(1)$.

In the case of the exchangeable correlation structure for within-community edges, hence $\lambda = 1$, the convergence rate in (2.12) decreases to the order of $\mathcal{O}_{N,M}\big\{ \exp(-\frac{rM}{min(r, \kappa_2 N)}) \big\}$, and therefore does not benefit from the increasing number of nodes. In this case, the consistency relies on accumulating independent sample networks. Theorem 2.1 also implies that the independent likelihood approach is unable to fully accumulate discriminative power from the increasing number of nodes when there exists dependency among within-community edges. Indeed, the convergence rate of the independent likelihood approach decreases in terms of network size N as the within-community correlation density λ increases. However, we show that the proposed approximate likelihood approach still benefits from increasing nodes size even under the exchangeable correlation structure among edges.

Theorem 2.2. *Under the regularity conditions (C1)-(C3), we establish the convergence rate of the estimator* z *using the proposed approximate likelihood approach. That is, for every* $t > 0$, $z \neq z^*$,

and $\lambda > 0$*,*

$$P_{Z*}\Big\{\frac{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z}=z;\Theta)}{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z}=z^*;\Theta)} > t\Big\} = \mathcal{O}\Big(\exp\Big\{-C_2\frac{r\lambda NM(c^*\eta_N+\lambda N^2)}{1+\rho\kappa_2 N\min(r,\kappa_2\lambda N)}\Big\}\Big), \tag{2.14}$$

where $N > O_N(\frac{1}{\lambda})$*,* $r = \|z - z^*\|_0$ *is the number of misclassified nodes up to the permutation labeling,* C_2 *is a positive constant,* ρ *is the largest within-community correlation, and* c^* *is defined in Theorem 2.1.*

Similarly, we characterize the convergence of the proposed method by the following corollary:

Corollary 2.2. *Under the same conditions given in Theorem 2.2, the proposed approximate likelihood approach leads to the following convergence rate, for every* $t > 0$

$$P_{Z*}\Big\{\sup_{\{z\neq z^*\}}\frac{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z}=z;\Theta)}{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z}=z^*;\Theta)} > t\Big\} = \mathcal{O}\Big(N\exp\Big\{-\frac{(c^*\eta_N+\lambda N^2)M}{N}\Big\}\Big). \tag{2.15}$$

Given $\lambda > 0$, for the same number of network M and node size N, the proposed approximate likelihood approach is able to achieve a faster convergence rate in (2.15) compared with (2.13) since the convergence rate in (2.14) has an additional term of $\lambda^2 N^3 M$ on the numerator compared to the convergence rate in (2.12). Specifically, the proposed approach is most superior under the exchangeable correlation structure ($\lambda = 1$), where the convergence rate of the independent likelihood is at the order of $\mathcal{O}_{N,M}\big\{\exp(-M/N)\big\}$, in contrast to the proposed convergence rate of $\mathcal{O}_{N,M}\big\{\exp(-NM)\big\}$. Intuitively, incorporating the correlation information increases the effective sample size of within-community edges. Under the sparsity assumption of higher-order correlation among edges, the proposed approach benefits from accumulating information on the second-order interactions among edges, while the independent likelihood approach only accumulates information from the first-order marginal mean of edges. It is noticeable that the marginal sparsity η_N affects the convergence rate of the membership estimator not only through marginal information but also through its constraints on the edge-wise correlation with intensity λ. In general, the number of edges correlated with other nodes decreases as sample networks become sparser in that $0 \leq \lambda \leq \mathcal{O}_N(\eta_N)$.

2.5.2 Computational Convergence for the Proposed Algorithm

In this subsection, we provide the computational convergence property of the proposed algorithm in Section 4. The main difference between the proposed method and the variational EM lies in the Bayes factor of (2.9) in the expectation step from Algorithm 1. If we replace $\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z})$ by the conditional independent likelihood $L_{ind}(\boldsymbol{Y}|\boldsymbol{Z})$ in (2.4) in the expectation step, the standard variational EM becomes a special case of Algorithm 1. Notice that [120] establishes computational convergence with the minimax rate of misclassification only when the within-community edges are independent. In addition, it assumes that the within-community marginal means are all the same, which is too restrictive in practice.

In the following, we establish the computational convergence for the proposed approximate likelihood. Specifically, we are able to show a faster convergence speed and a lower estimation bias compared to the existing one based on the independent likelihood in [120]. The following Theorem 2.3 also relaxes the homogeneous marginal mean assumption and allows the marginal means from within-community and between-community to be different. We denote the estimated memberships of nodes at the sth iteration as $\boldsymbol{\alpha^{(s)}} = (\alpha_1^{(s)}, \cdots, \alpha_N^{(s)})$ from Algorithm 1. In addition to the assumptions (C1-C3) in Section 5.1, we require two regularity conditions for the following theorems:

(C4). Suppose the distance between initial membership $\boldsymbol{\alpha}^{(0)}$ and true membership z^* is bounded: $\|\boldsymbol{\alpha}^{(0)} - z^*\|_1 \leq cN^{1-\phi}$, where $\phi \in (0, 1)$ is a constant.

A common issue for most EM-type algorithms including the one proposed is that they only guarantee convergence to a local optimum. If the likelihood function is unimodal, then the EM-type algorithm converges to the MLE as the unique global optimum. However, the proposed approximate likelihood is non-convex and multi-modal. Therefore, we assume that the initials are in the neighborhood of the MLE to ensure the convergence of the EM algorithm[15, 118]. Condition C4 is a common assumption to guarantee computational convergence for EM-type algorithms [120, 54, 57].

(C5). The estimated marginal mean $\hat{\mu}_{ql}$ has a bounded bias from the truth, i.e., $0 < \gamma_1 \leq \frac{\hat{\mu}_{ql}}{\mu_{ql}} \leq \gamma_2,\ q, l = 1 \cdots, K,$.

Theorem 2.3. *Under the regularity conditions (C1)-(C5) and given N is sufficiently large, we*

establish the convergence property of Algorithm 1 via incorporating correlation information. That is, with the correlation density $\frac{1}{\lambda} = o_N(N^{\frac{\phi}{2}})$*, as* M *and* N *increase with* $O_N(\frac{1}{\lambda}) < N$ *and* $M \leq o(N^{2-\phi/2})$*, then*

$$E\|\boldsymbol{\alpha}^{(s+1)} - z^*\|_1 \leq c_1 NK \exp\big\{ - c_2 \frac{\lambda(c^*\eta_N N + \lambda N^3)M}{1+\lambda N^2}\big\} + \frac{c_3\|\boldsymbol{\alpha}^s - \boldsymbol{z^*}\|_1}{\lambda NM}, \tag{2.16}$$

where c_1, c_2, c_3 *are positive constants, and* c^* *is defined in Theorem 2.1.*

In Theorem 2.3, the first term on the right hand side of the inequality represents the estimation bias which measures the discrepancy between the community structure and its realization. Although we do not show that the order of estimation bias in the first term achieves the minimax rate, there is a connection between our result and the minimax rate when $M = 1$. That is, our theorem implies that it has the same order as the minimax rate established for a single network case in Theorem 3 of [120].

The second term provides a decreasing rate of misclassification along each iteration. Theorem 2.3 indicates that the estimated memberships are closer to the true memberships compared to the previous iteration step at a rate of $\frac{1}{\lambda NM}$, where a larger sample size M or node size N contribute a faster convergence and a lower estimation bias. In general, Theorem 2.3 guarantees the convergence of the iterative algorithm even without incorporating correlation information, but improves the convergence rate and estimation bias when correlation information is incorporated.

Specifically, in contrast to the computational convergence rates in Theorem 3.1 of [120], our Theorem 2.3 shows that incorporating the correlation information enables us to reduce the estimation bias and accelerate the convergence rate. Specifically, if we consider the M sample networks generated independently from a SBM with the node size N, then the proposed approximate likelihood approach reduces the order of the estimation bias to $\mathcal{O}_{N,M}\big\{N\exp(-c'\lambda NM)\big\}$ in (2.16) as the dependency intensity λ increases and approaches to $\mathcal{O}_{N,M}\big\{N\exp(-cNM)\big\}$, which is equivalent to the minimax rate established in [120] treating edges as independent. Compared with the convergence rate of the membership estimator assuming conditional independence, incorporating within-community correlation accelerates the computational convergence rate from $\mathcal{O}_{N,M}(\frac{1}{\sqrt{MN}})$ [120] to $\mathcal{O}_{N,M}(\frac{1}{\lambda NM})$ in (2.16) when there is a sufficiently large number of correlated edges satisfying $\lambda > \frac{1}{\sqrt{MN}}$ within a community.

2.6 Numerical Studies

In this section, we conduct simulation studies to illustrate the performance of the proposed method on community detection in networks for dependent edges within-community. In particular, we compare our method to the existing variational EM method which assumes conditional independence among edges. Besides the comparison between the proposed method and independent likelihood method, we also conduct numerical comparisons between the propose method and existing multiple network community detection methods under different within-community dependency structure. Specifically, [70] proposes a spectral methods based on the optimal weighted average of multiple adjacent matrices (weighted average network), and the weighted average low-rank approximation (WALRA) which replaces an average of adjacent matrices by an average of low-rank approximation to each adjacent matrix. [113] proposes to jointly embed multiple adjacent matrices to a common subspace for clustering (joint embedding). [67] proposes an EM-based algorithm to recover community structure from the multiple noisy realizations of network (network denoising).

2.6.1 Study 1: Networks with Dependent Within-community Connectivity

In the first simulation study, we consider networks where edges within the same community are correlated and compare the performance of various methods under different network sample sizes with various magnitudes of marginal means for within-community and between-community.

Suppose the memberships of nodes $\boldsymbol{Z}^* = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n\}$ in the networks are given with K communities, where $\boldsymbol{Z}_i$ is a binary indicator vector corresponding to the membership of nodes i. Conditional on $\boldsymbol{Z}^*$, edges in each sample network are generated following the Bernoulli marginal distribution as in (2.1), where within-community edges follow an exchangeable correlation structure as in (2.5). Here we assume that between-community edges are independent from each other. The block-wise marginal means μ_{ql} $(q, l = 1, \cdots, K)$ are associated with edgewise covariates through (2.2). In addition, the edgewise covariates follow a uniform distribution, where within-communities covariates

$$x_{ij}^m \sim Unif(a_1, a_2) \text{ if } Z_{iq} = Z_{jq} = 1, \tag{2.17}$$

and between-community covariates

$$x_{ij}^m \sim Unif(b_1, b_2) \text{ if } Z_{iq} \neq Z_{jq}, q = 1, \cdots, K. \tag{2.18}$$

Although the probability of each edge is different, the edges within the same community share the same coefficient β_{ql} in (2.2). In the following simulation studies, we generate correlated unweighted edges through the R package "MultiOrd."

Specifically, the sample networks consist of 40 nodes split into two communities. In a balanced community network, each community has 20 nodes. In an unbalanced case, two communities are comprised of 10 and 30 nodes, respectively. We compare the performance under different sample sizes of networks with $M = 20,\ 40$ and 60, and different intensities of within-community dependency with correlation coefficient $\rho = 0,\ 0.3$ and 0.6.

To simulate a weak marginal signal case, we let the block-wise parameters be $\beta_{11} = 1,\ \beta_{22} = 1.5$ and $\beta_{12} = \beta_{21} = 0$. The means of within-community and between-community covariates are 0 with $a_1 = b_1 = -0.2$ and $a_2 = b_2 = 0.2$ in (2.17) and (2.18). Here, although the marginal mean of within-community edges is slightly larger than that of between-community edges on average due to the convexity of the logistic link function in (2.2), the marginal means of within-community edges and between-community edges are very close.

For a strong marginal signal case, the block-wise parameters are $\beta_{11} = 0.3,\ \beta_{22} = 0.6$ and $\beta_{12} = \beta_{21} = 0.2$. The within-community covariates are generated via (2.17) with $a_1 = 0.9$ and $a_2 = 1.1$, and between-community covariates are generated from (2.18) with $b_1 = -0.8$ and $b_2 = -0.6$. Note that there is a distinct gap between within-community and between-community marginal means, thus the marginal signal is more dominant for nodes within communities.

We use the Adjusted Rand Index (ARI) to measure the performance of clustering. The ARI takes a value between -1 and 1, where 1 represents a perfect matching of true memberships and predicted memberships of clustering, 0 indicates a random clustering and a negative value indicates that the agreement is less than the expectation from a random result. In the following simulations, we choose five fixed initial memberships of nodes in both balanced and unbalanced communities. These initials can be obtained from spectral clustering on sample networks. The Adjusted Rand Indices based on these chosen initials range between 0.30 to 0.34 under the unbalanced community

case and between 0.25 to 0.29 under the balanced community case, which are far from the true memberships.

We compare the performance of clustering and parameter estimation for the proposed method applying the second-order (Bahadur$_{\text{2nd}}$) and the fourth-order (Bahadur$_{\text{4th}}$) Bahadur approximation, and the variational EM (VEM) approach with only marginal information.

In Table 2.1 and Table 2.2, the proposed method with the second-order and fourth-order approximations outperform the variational EM in clustering. Specifically, under the weak marginal signal case in Table 1, the Adjusted Rand Index of the variational EM are 0.34 under different network sizes and correlation strengths, which are similar to the ones calculated by fixed initials. In addition, since the distributions of marginal means from within-community and between-community are similar, the variational EM marginal approach barely improves over the initial memberships as it only utilizes the marginal information. However, the proposed method with the second-order or fourth-order Bahadur representation improves on the ARI by about 280%, compared to the VEM when $\rho = 0.3$ and $\rho = 0.6$. In addition, the performance of the proposed method improves by $1 \sim 5\%$ as the number of sample networks increases from 20 to 60. Furthermore, incorporating the fourth-order interaction can slightly improve the accuracy of clustering.

We notice that when the correlation is as moderate as 0.3, the proposed method still achieves significant improvement over the variational EM and almost fully recovers the true memberships of clustering. We consider this as an intrinsic advantage of the proposed method in capturing the relatively weak dependency among edges to improve the clustering. This is because the proposed method not only captures pairwise dependency but also reflects connectivities among nodes within a community. That is, even a weak dependency among pairwise connectivities can lead to an accumulative information recovery of clustering.

Table 2.2 illustrates the clustering performance when the marginal signal is strong. In contrast to Table 2.1, the variational EM significantly improves on clustering because of the large discrepancy between the within-community marginal mean and the between-community marginal mean. Nevertheless, incorporating the correlation among within-community edges still improves the clustering accuracy by 20% to 26% under various sample sizes of networks and intensities of correlation. The clustering accuracy of the proposed method improves when either the sample size or the correlation increases. In general, stronger correlation and a larger sample size lead to better

performance when the marginal signal itself is strong.

In addition to clustering, we also provide estimation of the marginal parameters. Tables 2.3, 2.4 and 2.5 compare parameter estimation between the proposed method and the variational EM when the marginal signal is weak. For within-community parameters β_{11} and β_{22}, the estimation of the proposed method consistently reduces bias $30 \sim 99\%$ more than the variational method, except when $M = 20$ and $\rho = 0.6$. This is because the sample size $M = 20$ is not sufficiently large to offset the high variance among highly-correlated within-community edges. For the between-community parameter β_{12}, the estimation bias of the proposed method consistently decreases more than 80% compared to the VEM under all settings. Additionally, the standard errors of the proposed estimator decrease faster than the variational method as the sizes of networks increase.

For simulation settings, we generate the probability for within-community edges as 0.55 and for between-community edges is 0.50. We generate a mixture correlation structure to include both positive and negative pairwise correlations among edges within a community with equal proportions. The magnitude of pairwise edge correlation is $|\rho| = 0.5$ for all the correlation structure settings.

The results in Table 2.8 indicate that the propose method outperforms other competing methods for various correlation structure settings as most of competing methods only utilize the first-order marginal discrepancy among within-community and between community edges, but ignore the second-order discrepancy. The improvement from the proposed method incorporating the second-order discrepancy is more significant under the mixture correlation setting as the marginal discrepancy between intra-community and inter-community edges is less compared to other correlation structure such as the exchangeable or AR(1) with positive within-community correlations.

To demonstrate the practical feasibility of the proposed method on handling large networks, we further conduct simulations on the settings where each sample network consists of 500 nodes with two balanced communities. Specifically, within each community, there exists a subgroup of 100 nodes such that edges within the same subgroup are pairwisely correlated with the mixture correlation structure. To accelerate the convergence, the propose method adopts nodes' membership estimations from network denoising [67] as a warm start. The result provided in Table 2.9 indicates that the proposed method is able to improve classification accuracy on the nodes' memberships using a warm start initialization, and outperforms competing methods.

We also investigate the clustering performance of the independent likelihood and the proposed approximate likelihood approach given different within-community second-order correlation density λ in (**??**). The setting is similar to the weak marginal signal cases. Specifically, the sample networks contain two communities with identical pairwise within-community correlation $\rho = 0.6$. The sizes of the sample networks and nodes are $M = 40, N = 40$. The density λ increases from 0.01 to 1. The Adjusted Rand Index comparisons are illustrated in Figure 2.2. In general, the approximate likelihood approach has improving performance when the correlation connectivities among within-community edges increase, in contrast to the independent likelihood approach. Figure 2.2 shows that the true membership recovery using the approximate likelihood approach is high even when the second-order within-community correlation is relatively sparse ($\lambda = 0.05$), while the independent likelihood approach performs poorly with a constant ARI regardless of λ . This finding supports Theorem 2.1 and 2.2 in that the proposed method produces an accelerated decay in misclassification rate as λ increases.

2.6.2 Study 2: Networks with Additional Dependence between Different Communities

In Study 2, we also investigate whether the proposed method holds for a more general dependency structure among edges from different communities, for example, correlation among edges between different communities

$$corr(Y^m_{i_1j_1}, Y^m_{i_2j_2}) = \tilde{\rho}, \text{ given } z_{i_1} = z_{j_1} = q,\ z_{i_2} = z_{j_2} = l,\ q \neq l, \tag{2.19}$$

where $\tilde{\rho} \leq \rho_q$ in (2.5) in general. While (2.5) characterizes the concordance of edges within a community, (2.19) also captures the heterogeneity of sample networks. The heterogeneity of multi-layer networks is common in community detection.

In this simulation, we demonstrate that the proposed method is still robust when there is heterogeneity of connectivities among sample networks. To simulate the dependency among inter-community connectivity, we split M sample networks into 10 groups. Within each group, we add

the random effects γ_k to the within-community marginal means:

$$\mu_{qq}^m = \frac{exp(\beta_{ql} x_{ij}^m)}{1 + exp(\beta_{ql} x_{ij}^m)} + \gamma_k,\ M\frac{k-1}{10} \leq m \leq M\frac{k}{10},$$

where $\gamma_k \sim N(0, \sigma^2)$, $k = 1, \cdots, 10$, $m = 1, \cdots, M$, and $q = 1, \ldots, K$. The variance σ of the random effect γ_k captures the intensity of dependency among inter-community connectivities, which increases as σ increases. We set $\sigma = 0.5$ to represent a weak inter-community dependency and $\sigma = 1.5$ for a strong inter-community dependency, while the other settings remain the same as in simulation Study 1. Our primary interest is to compare clustering performance between the proposed method and the variational method under the weak marginal signal case.

Tables 2.6 and 2.7 illustrate the clustering performance between the variational method and the proposed method under balanced and unbalanced community sizes, respectively. When the within-community correlation is moderate at 0.3, the proposed method improves the clustering accuracy by 170% to 257% for various network sizes and σ. For strong correlation $\rho = 0.6$, the improvement is between 210% to 257%. In particular, the proposed method has better performance when the networks have strong intra-community correlation and large sample sizes under both weak and strong inter-community correlation cases. In addition, using the fourth-order Bahadur representation improves the accuracy by 6% and 14% when $\sigma = 0.5$ and $\sigma = 1.5$ compared to the second-order Bahadur representation, indicating that the higher-order method still enhances the clustering outcome under the misspecified model. It is interesting to note that the performance of the proposed method decreases by 5% to 15% when the inter-community correlation is strong and the number of networks is small, compared to the same setting with weak inter-community correlation. However, the performances under both weak or strong inter-community correlation are similar when the sample size of networks increases. In conclusion, the proposed method is robust against misspecified dependency structure when the sample size increases.

2.7 Real Data Example

In this section, we apply the proposed method to the 2010 Worldwide Food Import/Export Network dataset [38] from `https://github.com/CompNet/MultiplexCentrality/tree/master/`

`data/FAO_Multiplex_Trade`. We create 364 networks among 214 countries with a total of 318,346 edges, where each network captures the trading connections of a specific food product among countries.

The primary goal of the study is to identify the common trading communities among different countries shared by food and agricultural product networks. The phenomenon of common community structures for international food-trade multi-networks has been recently studied and supported by [110, 69, 17, 9]. In general, the community structure in trade networks of food products is highly influenced by factors of geographical, climatic, socio-economic and political relations among different countries.

One significant feature of these networks is that the average empirical correlation of the pairwise connection among trading countries is 0.29. Therefore, the SBM based on the conditional independent assumption among edges could possibly lead to a biased network clustering of countries.

We first preprocess the data to select nodes corresponding to the trading countries which are most relevant, the number of communities and the initial memberships of countries. Note that several major countries dominate the world economy and lead a high number of trading connectivities, while the other countries with limited agricultural product categories have fewer trading connections with other countries for specific product networks. Here we focus on the partial trading networks consisting of major countries whose corresponding degrees of nodes are larger than 9, which results in 51 countries with major economic impact in the world, such as the United States, mainland China, Japan and some European countries. The average empirical correlation of the trading connections among these countries is 0.22, indicating that the connectivity dependency should be considered in clustering these countries' trading networks.

In general, there are two major procedures to select the number of communities. First, we can perform the Louvain method for community detection on each individual trading network to obtain the number of communities which maximizes the modularity and the size of the largest community. Next we take the average of the number of communities on networks whose number of communities is smaller than 10 and whose largest community size is larger than 14. This procedure removes the 18% of the product trading networks whose countries are commercially isolated from other countries, as our goal is to detect the commercial communities among the countries which are more connected with other countries. After preprocessing, the average number of communities

is 4.9 and we set it to be 4, and there are 296 sample networks remaining in the following analysis.

Table 2.10 and Figure 2.2 provide the estimated agricultural products trading communities among 51 countries based on the independent likelihood model using variational EM and the proposed method. Table 2.11 provides the estimation of international communities from the competing methods mentioned in Section 2.6. In the following, we focus on the comparison of clustering results between the independent likelihood method and the proposed method as the results from comparisons between the proposed method and other competing methods could be inferred similarly.

For the proposed method, we implement the fourth-order Bahadur approximation since it can better capture high-order within-community connectivity dependency. Table 2.10 presents the clustering outcome among countries according to the variational method and the proposed method. The countries in the same community under the variational method are marked with the same color, while the newly formed communities based on the proposed method are illustrated on the right sides of Table 2.10 and Figure 2.2. In general, the Adjusted Rand Index for clustering between the variational method and the proposed method is 0.43, indicating that the communities detected by the two methods are quite different. The clustering results from the proposed method incorporating within-community dependency are more interpretable compared to the variational EM using only marginal information.

In particular, the proposed method identifies communities 1 and 2 (red and cyan color communities on the right panel of Figure 2.2) which are highly associated with their geographical and climate environments. However, these features are not detected by the variational method. For example, community 1 with the cyan color on the left of Figure 2.2 based on the variational method mainly consists of two types of countries: one group comprises Nordic and Eastern European countries, and the other group consists of countries in Latin American and Africa. In contrast, the proposed method clusters countries from geographically neighboring countries in east Europe, including Austria, Poland and Romania which are clustered with other communities by the variational method. Community 2 with blue color on the left of Figure 2.2 based on the variational method contains northern countries such as Canada as well as tropical countries. However, the proposed method identifies community 2 with tropical coastal countries and Arabian Peninsula countries, which provides more meaningful community clusters compared to the variational EM

method.

The variational method and proposed method detect the same third community with orange color in Figure 2.2 which contains 7 major countries from the European Union: Belgium, France, Germany, Italy, Netherlands, Spain and the UK.

The fourth community from the variational method colored with red on the left of Figure 2.2 consists of 11 Eastern European countries, and all are categorized in community 1 from the proposed method. Community 4 with blue color on the right of Figure 2.2 in the proposed method includes countries with large populations or more developed agricultural product trading, such as mainland China, U.S.A, India and Japan.

Similar to the independent likelihood approach, the clustering of countries based on other competing methods (e.g., weighted average network, WALRA, and joint embedding) do not show clear intrinsic patterns or similarity among nations within the same community. In contrast, the proposed method groups countries based on geographical and climatic similarity. In particular, geographical distance can be differentiable in terms of the likelihood function on clustering countries for the same trading community across different products, and this finding is also supported by [110].

In terms of parameter estimation, the average probability of having trading connections for communities 1 and 2 based on the variational method are 0.21 and 0.52, respectively. For the proposed method, the estimated correlations of connectivities within communities 1 and 2 are both 0.22, and the corresponding average within-communities connection rates are 0.28 and 0.22, respectively. The relatively low connection rates and correlations may be related to the low diversity and high overlaps of product categories due to more restrictive geographical and climate environments.

For community 3, the corresponding estimated marginal parameters β_{33} from the proposed method and the variational method are 2.58 and 2.00 respectively, both of which indicate that the trading connection rate within European Union communities is greater than 88% on average. This strong marginal signal of within-community connection explains that the additional correlation information is less influential in clustering. Additionally, the estimated correlation within the third community is 0.58, implying a high connection rate within-community. For community 4, the corresponding average connection rate is 0.49 based on the variational method, and the estimated within-community average connection rate and the correlation are 0.61 and 0.27, respectively. This is because community 4 involves large population countries with more frequent trading on product

categories due to their higher food diversity than other countries.

2.8 Discussion

In this chapter, we propose a new community detection method for networks incorporating the underlying dependency structure among connectivities. To model the correlation without specifying a joint likelihood for correlated edges, we construct an approximate likelihood based on the Bahadur representation which decomposes a joint distribution into a marginal term and high-order interaction terms. The proposed method provides flexible modeling on the correlation structure which can be specified through the interaction term in the approximate likelihood.

In theory, we establish the consistency of the nodes' membership estimator based on the proposed approximate likelihood and show that it achieves a faster convergence than the independent method. In addition, we show that the proposed iterative algorithm possesses desirable convergence properties. In particular, we show that the proposed approximate approach can achieve a faster computational convergence and a lower clustering bias compared to the variational EM algorithm. Furthermore, we show that the variational EM algorithm is a special case of our algorithm under the conditional independent model, which confirms that incorporating correlation information improves the accuracy for community detection.

Our numeric studies indicate that incorporating the within-community correlation among edges can improve the clustering performance compared to the marginal model, even under a moderately misspecified model on inter-community dependency. The improvement of community detection is more significant when the marginal signal is weak, which is less informative for distinguishing between within-community and between-community networks. In addition, the proposed method enables us to achieve more accurate parameter estimation.

In this chapter, we only consider incorporating the within-community dependency. It would be worthy of further research to investigate more generalized dependency structures to include between-community dependency as well.

2.9 Figures and Tables

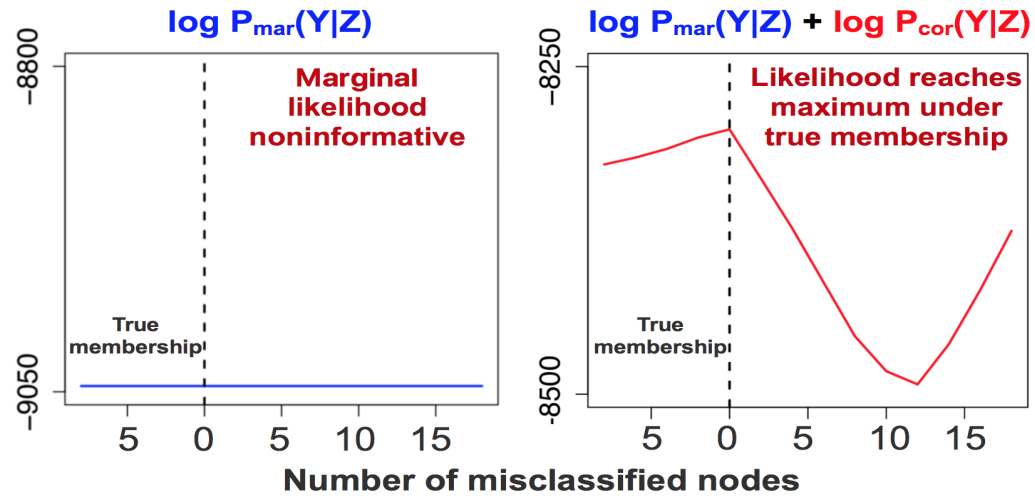


Figure 2.1: Likelihood of multiple networks with 30 nodes from two communities. *Left*: Traditional SBM likelihood. *Right*: The proposed pseudolikelihood incorporating correlation information.

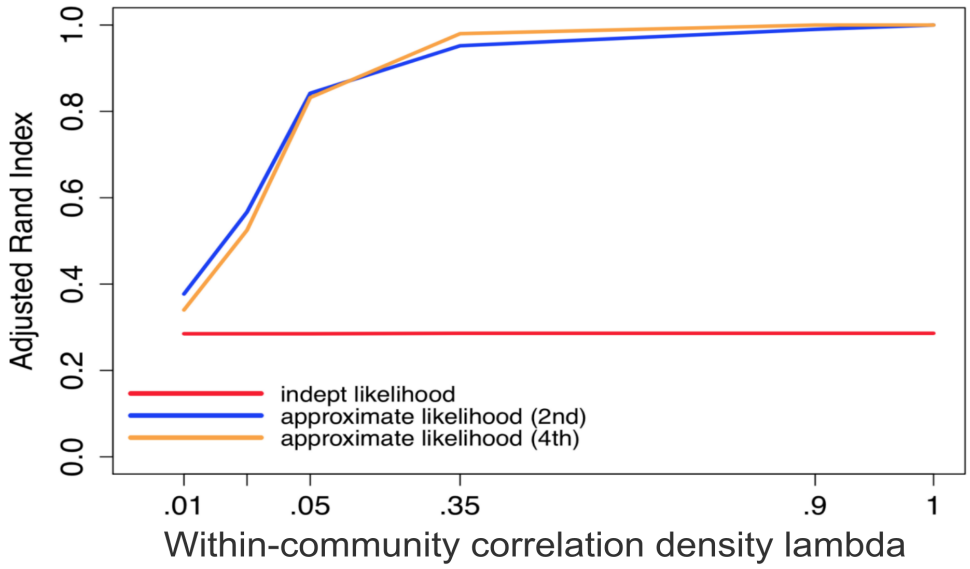


Figure 2.2: Clustering performance comparisons between independent likelihood and the proposed approximate likelihood approach incorporating the second-order and fourth-order correlations.

Table 2.1: Adjusted Rand Index between estimated membership and true membership for networks with two communities and weak marginal signal averaging on 50 replicates.

		Unbalanced community			**Balanced community**		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0$	VEM	0.38	0.41	0.48	0.31	0.28	0.28
	$\text{Bahadur}_{\text{2nd}}$	0.36	0.41	0.47	0.32	0.29	0.29
	$\text{Bahadur}_{\text{4th}}$	0.35	0.37	0.47	0.30	0.29	0.30
$\rho = 0.3$	VEM	0.34	0.34	0.34	0.28	0.28	0.28
	$\text{Bahadur}_{\text{2nd}}$	0.94	0.98	0.99	0.96	0.99	1.00
	$\text{Bahadur}_{\text{4th}}$	0.96	0.99	1.00	0.99	0.99	1.00
$\rho = 0.6$	VEM	0.34	0.34	0.34	0.29	0.28	0.28
	$\text{Bahadur}_{\text{2nd}}$	0.96	0.99	0.99	0.97	1.00	1.00
	$\text{Bahadur}_{\text{4th}}$	0.99	1.00	1.00	0.99	1.00	1.00

Table 2.2: Adjusted Rand Index between estimated membership and true membership for networks with two communities and strong marginal signal averaging on 50 replicates.

		Unbalanced community			**Balanced community**		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0$	VEM	0.78	0.92	0.98	0.76	0.90	0.97
	$\text{Bahadur}_{\text{2nd}}$	0.73	0.91	0.97	0.77	0.92	0.98
	$\text{Bahadur}_{\text{4th}}$	0.69	0.86	0.95	0.72	0.92	0.98
$\rho = 0.3$	VEM	0.78	0.81	0.83	0.68	0.79	0.84
	$\text{Bahadur}_{\text{2nd}}$	0.99	0.99	1.00	0.98	1.00	1.00
	$\text{Bahadur}_{\text{4th}}$	0.99	0.99	1.00	0.99	1.00	1.00
$\rho = 0.6$	VEM	0.78	0.89	0.83	0.84	0.92	0.88
	$\text{Bahadur}_{\text{2nd}}$	0.99	1.00	1.00	0.99	1.00	1.00
	$\text{Bahadur}_{\text{4th}}$	0.99	1.00	1.00	0.99	1.00	1.00

Table 2.3: Estimation of within-community parameter $\beta_{11} = 1$ for networks with two communities and weak marginal signal.

		Unbalanced community			**Balanced community**		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0$	VEM	$0.56_{0.42}$	$0.59_{0.29}$	$0.58_{0.20}$	$0.64_{0.32}$	$0.57_{0.16}$	$0.64_{0.18}$
	Bahadur$_{2nd}$	$0.57_{0.42}$	$0.58_{0.30}$	$0.57_{0.21}$	$0.61_{0.28}$	$0.57_{0.16}$	$0.66_{0.20}$
	Bahadur$_{4th}$	$0.52_{0.42}$	$0.55_{0.28}$	$0.57_{0.19}$	$0.58_{0.27}$	$0.58_{0.18}$	$0.65_{0.19}$
$\rho = 0.3$	VEM	$0.49_{0.30}$	$0.50_{0.17}$	$0.52_{0.14}$	$0.58_{0.24}$	$0.58_{0.18}$	$0.59_{0.12}$
	Bahadur$_{2nd}$	$0.81_{0.48}$	$0.84_{0.32}$	$0.89_{0.27}$	$0.95_{0.24}$	$0.93_{0.16}$	$0.92_{0.14}$
	Bahadur$_{4th}$	$0.85_{0.47}$	$0.83_{0.31}$	$0.89_{0.27}$	$0.96_{0.24}$	$0.93_{0.16}$	$0.93_{0.14}$
$\rho = 0.6$	VEM	$0.56_{0.22}$	$0.54_{0.20}$	$0.52_{0.15}$	$0.61_{0.27}$	$0.61_{0.16}$	$0.60_{0.14}$
	Bahadur$_{2nd}$	$1.01_{0.42}$	$1.04_{0.35}$	$1.00_{0.29}$	$0.95_{0.31}$	$1.00_{0.19}$	$0.96_{0.15}$
	Bahadur$_{4th}$	$0.99_{0.25}$	$1.05_{0.15}$	$1.01_{0.13}$	$0.97_{0.31}$	$1.01_{0.19}$	$0.97_{0.16}$

Table 2.4: Estimation of within-community parameter $\beta_{22} = 1.5$ for networks with two communities and weak marginal signal.

		Unbalanced community			**Balanced community**		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0$	VEM	$1.43_{0.43}$	$1.42_{0.34}$	$1.45_{0.26}$	$1.18_{0.40}$	$0.94_{0.16}$	$0.94_{0.15}$
	Bahadur$_{2nd}$	$1.50_{0.39}$	$1.49_{0.31}$	$1.45_{0.25}$	$1.21_{0.42}$	$0.93_{0.21}$	$0.97_{0.22}$
	Bahadur$_{4th}$	$1.56_{0.37}$	$1.49_{0.30}$	$1.46_{0.23}$	$1.19_{0.47}$	$0.94_{0.24}$	$0.96_{0.22}$
$\rho = 0.3$	VEM	$1.31_{0.23}$	$1.40_{0.11}$	$1.37_{0.11}$	$1.05_{0.21}$	$0.92_{0.16}$	$0.92_{0.16}$
	Bahadur$_{2nd}$	$1.56_{0.19}$	$1.50_{0.10}$	$1.49_{0.09}$	$1.48_{0.22}$	$1.45_{0.19}$	$1.44_{0.14}$
	Bahadur$_{4th}$	$1.55_{0.19}$	$1.50_{0.09}$	$1.49_{0.09}$	$1.48_{0.22}$	$1.45_{0.19}$	$1.45_{0.14}$
$\rho = 0.6$	VEM	$1.46_{0.16}$	$1.43_{0.16}$	$1.38_{0.13}$	$1.16_{0.21}$	$1.09_{0.21}$	$1.06_{0.22}$
	Bahadur$_{2nd}$	$1.73_{0.29}$	$1.60_{0.15}$	$1.52_{0.12}$	$1.73_{0.28}$	$1.60_{0.29}$	$1.64_{0.15}$
	Bahadur$_{4th}$	$1.69_{0.25}$	$1.60_{0.15}$	$1.52_{0.13}$	$1.73_{0.26}$	$1.61_{0.29}$	$1.64_{0.15}$

Table 2.5: Estimation of within-community parameter $\beta_{12} = 0$ for networks with two communities and weak marginal signal.

		Unbalanced community			**Balanced community**		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0$	VEM	$0.52_{0.35}$	$0.57_{0.24}$	$0.47_{0.22}$	$0.22_{0.31}$	$0.39_{0.14}$	$0.41_{0.11}$
	Bahadur$_{\text{2nd}}$	$0.51_{0.32}$	$0.58_{0.23}$	$0.48_{0.21}$	$0.23_{0.30}$	$0.41_{0.16}$	$0.39_{0.15}$
	Bahadur$_{\text{4th}}$	$0.51_{0.29}$	$0.63_{0.22}$	$0.48_{0.20}$	$0.25_{0.28}$	$0.40_{0.17}$	$0.41_{0.13}$
$\rho = 0.3$	VEM	$0.68_{0.24}$	$0.68_{0.13}$	$0.69_{0.10}$	$0.42_{0.14}$	$0.35_{0.12}$	$0.40_{0.10}$
	Bahadur$_{\text{2nd}}$	$-0.02_{0.25}$	$0.00_{0.15}$	$0.00_{0.11}$	$0.03_{0.20}$	$-0.05_{0.16}$	$-0.02_{0.12}$
	Bahadur$_{\text{4th}}$	$-0.02_{0.24}$	$0.00_{0.14}$	$0.00_{0.11}$	$0.03_{0.18}$	$-0.06_{0.16}$	$0.03_{0.12}$
$\rho = 0.6$	VEM	$0.72_{0.17}$	$0.71_{0.11}$	$0.70_{0.09}$	$0.41_{0.18}$	$0.45_{0.11}$	$0.48_{0.11}$
	Bahadur$_{\text{2nd}}$	$-0.05_{0.17}$	$-0.03_{0.13}$	$0.02_{0.11}$	$0.00_{0.19}$	$0.01_{0.12}$	$0.03_{0.12}$
	Bahadur$_{\text{4th}}$	$-0.04_{0.17}$	$-0.03_{0.13}$	$-0.02_{0.11}$	$-0.02_{0.18}$	$0.00_{0.12}$	$0.03_{0.11}$

Table 2.6: Performance comparison given misspecified inter-community correlation with balanced community and weak marginal signal averaging on 50 replicates.

		$\sigma = 0.5$			$\sigma = 1.5$		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0.3$	VEM	0.28	0.28	0.29	0.28	0.28	0.29
	Bahadur$_{\text{2nd}}$	0.90	0.99	1.00	0.76	0.99	0.99
	Bahadur$_{\text{4th}}$	0.96	1.00	1.00	0.87	0.98	1.00
$\rho = 0.6$	VEM	0.28	0.28	0.29	0.28	0.28	0.29
	Bahadur$_{\text{2nd}}$	0.94	0.99	1.00	0.87	0.99	1.00
	Bahadur$_{\text{4th}}$	0.99	1.00	1.00	0.94	0.99	1.00

Table 2.7: Performance comparison given misspecified inter-community correlation with unbalanced community and weak marginal signal averaging on 50 replicates.

		$\sigma = 0.5$			$\sigma = 1.5$		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0.3$	VEM	0.32	0.33	0.33	0.33	0.33	0.33
	$\text{Bahadur}_{\text{2nd}}$	0.89	0.98	0.99	0.89	0.95	0.97
	$\text{Bahadur}_{\text{4th}}$	0.95	0.99	0.99	0.93	0.94	0.94
$\rho = 0.6$	VEM	0.34	0.33	0.34	0.33	0.33	0.33
	$\text{Bahadur}_{\text{2nd}}$	0.91	0.96	0.98	0.91	0.95	0.94
	$\text{Bahadur}_{\text{4th}}$	0.95	0.96	0.97	0.92	0.93	0.92

Table 2.8: Performance comparison given weak marginal signal on the 40×40 networks with two communities.

	Exchange		**AR(1)**		**Mixture**	
Sample size	M =30	M = 50	M = 30	M = 50	M = 30	M = 50
Proposed method	0.977	1	0.951	1	0.955	1
Weighted average network	0.104	0.255	0.506	0.917	-0.01	0.02
Weighted average low-rank approx	1	1	0.218	0.148	0.027	0.026
Joint embedding	0.32	0.66	0.64	0.80	0.15	0.17
Network denoising	-0.002	0.09	0.91	0.99	0.367	0.637

Table 2.9: Performance comparison given weak marginal signal on the 80 sample networks where each has 500 nodes.

Proposed method	**Weighted ave network**	**WALRA**	**Network denoising**
0.64	0.13	0.08	0.52

Table 2.10: Clustering of nations in the agricultural products trading networks for 4 communities.

	VEM	**Bahadur$_{4th}$**
Community 1	Brazil, Denmark, Finland, Ireland Lebanon, Russia, Sweden, Switzerland Turkey, Ukraine, Argentina, Israel Mexico, Norway, Portugal, Chile South Africa, Qatar	Austria, Denmark, Finland, Ireland, Poland Russia, Sweden, Switzerland, Turkey Bulgaria, Croatia, Czech, Greece, Hungary Israel,Lithuania, Norway, Portugal Romania, Slovakia, Slovenia, Ukraine
Community 2	Australia, Canada, Hong Kong, Mainland Taiwan, India, Indonesia, Malaysia Japan, Philippines, Korea, Singapore Thailand, U.S.A, New Zealand	Brazil, Hong Kong, Taiwan, Indonesia Lebanon, Philippines, Korea, Argentina Mexico, Chile, New Zealand, Qatar South Africa
Community 3	Belgium, France, Germany, Italy Netherlands, Spain, United Kingdom	Belgium, France, Germany, Italy Netherlands, Spain, United Kingdom
Community 4	Austria, Poland, Bulgaria, Croatia Czech, Greece, Hungary, Lithuania Romania, Slovakia, Slovenia	Australia, Canada, Mainland, India Japan, Malaysia, Singapore, U.S.A Thailand

Table 2.11: Clustering of nations for the agricultural products trading networks based on competing methods

	Weighted average network	WALRA	Joint embedding
Com 1	Brazil, Denmark, Finland, Ireland, Lebanon, Russian, Sweden, Switzerland Turkey, Ukraine, Argentina, Israel Lithuania, Mexico, Norway, Portugal Chile, South Africa, Qatar	Austria, Denmark, Finland, Romania, Ireland, Lebanon, Poland, Slovakia Russian, Sweden, Switzerland Turkey, Ukraine, Bulgaria, Croatia Israel, Lithuania, Norway, Portugal Czech, Greece, Slovenia, Hungary	Australia, Brazil, Hong Kong, Taiwan, Indonesia, Japan, Lebanon Malaysia, Philippines, Korea Thailand, Argentina, Mexico Chile, New Zealand, Canada Singapore, South Africa, Qatar
Com 2	Taiwan, Canada, Mainland, New Zealand Hong Kong, Australia, Singapore India, Indonesia, Thailand, Japan" Malaysia, Philippines, Korea, U.S.A	Mainland	Mainland, India, U.S.A
Com 3	Belgium, France, Germany, Italy Netherlands, Spain, United Kingdom	Belgium, France, Germany, Italy Netherlands, Spain, United Kingdom	Belgium, France, Germany, Italy Netherlands, Spain, United Kingdom
Com 4	Austria, Poland, Bulgaria, Croatia Czech, Greece, Hungary Romania, Slovakia, Slovenia	Australia, Brazil, Canada, Mexico Hong Kong, Taiwan, India, Chile Indonesia, New Zealand, Japan Malaysia, Philippines, Korea, Qatar Singapore, Thailand, U.S.A Argentina, South Africa	Austria, Denmark, Finland, Ireland Poland, Russian, Sweden, Switzerland Turkey, Ukraine, Bulgaria, Croatia Czech, Greece, Hungary, Israel Lithuania, Norway, Portugal Romania, Slovakia, Slovenia

2.10 Notation and Proofs

2.10.1 Notation

In the following, we denote the membership of node as random variable $z_i, i = 1, \cdots, N$. Then $\boldsymbol{Z} = \{z_1, z_2, \cdots, z_N\}$. Accordingly, we define the true membership of nodes as $z_i^* \in \{1, 2, \cdots, K\}$, $i = 1, \cdots, N$ and $z^* = \{z_1^*, z_2^*, \cdots, z_N^*\}$. We denote $P^*(\cdot) = P(\cdot|\boldsymbol{Z} = z^*)$ as the conditional probability of observed networks given the true nodes' membership z^*. The number of misclassified nodes is denoted as r such that $\|z - z^*\|_0 = r$ for $z \neq z^*$. Define the t-th sample network as $\boldsymbol{Y}^t = (Y_{ij}^t)_{N\times N}$ and t-th sample network standardized by $\hat{\mu}_{aa}$ as $\hat{\boldsymbol{Y}}^{t,a} = (\hat{Y}_{ij}^{t,a})_{N\times N}$ where $\hat{Y}_{ij}^{t,a} = \frac{Y_{ij}^t - \hat{\mu}_{aa}}{\sqrt{\hat{\mu}_{aa}(1-\hat{\mu}_{aa})}}, a = 1, \cdots, K,\ t = 1, \cdots, M$. We further define the s-th column of $\hat{\boldsymbol{Y}}^{t,a}$ as $\hat{Y}_{\cdot s}^{t,a}$. ρ_{ijuv} denotes pairwise correlation between two edges Y_{ij}^t and Y_{uv}^t. Given the empirical estimation $\hat{\rho}_{ijuv} = \rho_{ijuv}$ almost sure as M increase, we assume $\{\rho_{ijuv}\}$ are known in the following proofs.

Denote $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_N)$ as the estimated probability of nodes' memberships. Specifically, let $\alpha_i = (\alpha_{i1}, \cdots, \alpha_{iK})_{1\times K}$ be the probability of nodes i belonging to each community where $\sum_{q=1}^K \alpha_{iq} = 1,\ i = 1, \cdots, N$. For simplicity of notation, if the subscripts indicate the community then $\alpha_q = (\alpha_{1q}, \cdots, \alpha_{Nq})_{1\times N}$ represents the probability of each node belonging to community q, where $q = 1, \cdots, K$. Similarly, $z_q^* = \{z_{1q}^*, z_{2q}^*, \cdots, z_{Nq}^*\}$ is a binary vector indicating nodes whose true membership belongs to community $q, q = 1, \cdots, K$. Let $vec(\cdot)$ stand for the operation of vectorizing a matrix into a column.

The following lemma is introduced as the technical steps in the proofs of Theorem 2.1, Theorem 2.2 and Theorem 2.3. The proofs of Lemma 1 is provided in the supplemental material.

lemma 2.1. *Consider function* $f_1(x) = \sqrt{\left\{x \log \frac{\mu_{z_i z_j}}{\mu_{z_i^* z_j^*}} + (1-x)\log \frac{1-\mu_{z_i z_j}}{1-\mu_{z_i^* z_j^*}}\right\}_+}$ *and denote*

$$X_t^+ = \{f_1(Y_{12}^t), f_1(Y_{13}^t), \cdots, f_1(Y_{N-1,N}^t)\}$$

where $\{Y_{ij}^t\}_{N\times N}$ *are generated through the stochastic block model in section 3.1 and satisfy condition C1, C2 and C3. Define the covariance matrix of* X_t^+ *as* Σ_1. *Then* X_t^+ *is a subgaussian vector,*

i.e.,

$$L = inf\{\alpha \geq 0 : E(\exp(\langle z, X_t^+ - E(X_t^+)\rangle)) \leq \exp\{\alpha^2 \langle \Sigma_1 z, \ \ z\rangle\}/2, z \in R^{N(N-1)/2}\} \leq C$$

for some positive constant C.

Proof: recall that X_t^+ is a binary vector. For any random vector z such that $dim(z) = dim(X_t^+)$, consider random vectors $\varepsilon = \Sigma_1^{1/2} z, U_t = \Sigma_1^{-1/2}\{X_t^+ - E(X_t^+)\}$. Therefore,

$$\mathrm{Var}(U_t) = \Sigma_1^{-1/2}\Sigma_1\Sigma_1^{-1/2} = I.$$

Given each element in U_t is bounded such that $|(U_t)_i| \leq C_1$ and $E((U_t)_i) = 0$,$1 \leq i \leq \frac{n(n-1)}{2}$, we have

$$\begin{aligned}
&E\{\exp\big(\langle z, X_t^+ - E(X_t^+)\rangle\big)\} \\
=&E\big\{\exp\big(\langle \Sigma_1^{1/2} z, \Sigma_1^{-1/2}(X_t^+ - E(X_t^+))\rangle\big)\big\} = E\big\{\exp\big(\langle \epsilon, U_t\rangle\big)\big\} \\
=& E\{\prod_{i=1} \exp\big(\epsilon_i (U_t)_i\big)\} = E\{E(E(E(V_1)V_2)V_3)\cdots V_{n(n-1)/2}\},
\end{aligned}$$

where

$$\begin{aligned}
V_1 &= E\{\exp\big(\epsilon_1 (U_t)_1\big)|(U_t)_2, \cdots, (U_t)_{n(n-1)/2}\}, \\
V_2 &= E\{\exp\big(\epsilon_2 (U_t)_2\big)|(U_t)_3, \cdots, (U_t)_{n(n-1)/2}\}, \\
&\vdots \\
V_{n(n-1)/2} &= E\{\exp\big(\epsilon_{n(n-1)/2} (U_t)_{n(n-1)/2}\big)\}.
\end{aligned}$$

According to the Hoeffding's lemma, we have

$$V_i \leq \exp\{\frac{\epsilon_i^2 C_1^2}{2}\}, \ i = 1, \cdots, n(n-1)/2.$$

Therefore,

$$E\{\exp\big(\langle z, X_t^+ - E(X_t^+)\rangle\big)\} \leq \prod_{i=1} \exp\{\frac{\epsilon_i^2 C_1^2}{2}\} = \exp\{\frac{C_1^2}{2}\langle \epsilon, \epsilon\rangle\}$$
$$= \exp\{\frac{C_1^2}{2}\langle \Sigma_1 z, z\rangle\}.$$

Therefore, X_t^+ is a subgaussian random vector. In addition, denote L as subgaussian norm of X_t^+ such that

$$L = inf\{\alpha \geq 0 : E(\exp(\langle z, X_t^+ - E(X_t^+)\rangle)) \leq \exp\{\alpha^2\langle \Sigma_1 z, z\rangle\}/2\}.$$

Then we have $L \leq \frac{C_1^2}{2}$.

2.10.2 Proof of Theorem 2.1

Given the independent model in (2.4), we can simplify the likelihood ratio between a random membership z and the true membership z^* as

$$\log \frac{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z})}{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z^*})} = \frac{1}{M}\sum_{t=1}^{M}\sum_{i<j}\big\{Y_{ij}^t \log\frac{\mu_{z_i z_j}}{\mu_{z_i^* z_j^*}} + (1-Y_{ij}^t)\log\frac{1-\mu_{z_i z_j}}{1-\mu_{z_i^* z_j^*}}\big\}. \tag{2.20}$$

We define two transformation functions $f_1(x)$ and $f_2(x)$ as:

$$f_1(x) = \sqrt{\big\{x\log\frac{\mu_{z_i z_j}}{\mu_{z_i^* z_j^*}} + (1-x)\log\frac{1-\mu_{z_i z_j}}{1-\mu_{z_i^* z_j^*}}\big\}_+},$$
$$f_2(x) = \sqrt{\big\{x\log\frac{\mu_{z_i z_j}}{\mu_{z_i^* z_j^*}} + (1-x)\log\frac{1-\mu_{z_i z_j}}{1-\mu_{z_i^* z_j^*}}\big\}_-}.$$

where $\{\}_+$ and $\{\}_-$ are positive part and negative part of a random variable. The previous summation can be decomposed as positive part and negative part:

$$\log \frac{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z})}{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z^*})} = \frac{1}{M}\sum_{t=1}^{M}\sum_{i<j}\{f_1^2(Y_{ij}^t) - f_2^2(Y_{ij}^t)\}.$$

Define the vectorized edges in the t th sample network as:

$$X_t^+ = \{f_1(Y_{12}^t), f_1(Y_{13}^t), \cdots, f_1(Y_{N-1,N}^t)\}, X_t^- = \{f_2(Y_{12}^t), f_2(Y_{13}^t), \cdots, f_2(Y_{N-1,N}^t)\}. \quad (2.21)$$

Note that each element in X_t^+ or X_t^- is a bounded binary random variable. In addition, as $f_1(Y_{ij}^t)$ or $f_2(Y_{ij}^t)$ only rescale Y_{ij}^t then they preserve the within-community correlation among Y_{ij}^t. Then we consider the following quadratic forms

$$Q_1 = \sum_{t=1}^{M} \langle X_t^+, X_t^+ \rangle, Q_2 = \sum_{t=1}^{M} \langle X_t^-, X_t^- \rangle.$$

such that

$$\log \frac{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z})}{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}^*)} = \frac{1}{M}(Q_1 - Q_2) \text{ and } E(\log \frac{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z})}{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}^*)}) = \frac{1}{M}(EQ_1 - EQ_2).$$

Denote the centralized version quadratic forms Q_1 and Q_2 as $\mathcal{Q}_1$ and $\mathcal{Q}_2$ such that

$$\mathcal{Q}_1 = \sum_{t=1}^{M} \langle X_t^+ - E(X_t^+), X_t^+ - E(X_t^+) \rangle, \mathcal{Q}_2 = \sum_{t=1}^{M} \langle X_t^- - E(X_t^-), X_t^- - E(X_t^-) \rangle.$$

Denote the following quadratic difference as:

$$\Delta(Q_1, \mathcal{Q}_1) := (Q_1 - E(Q_1)) - (\mathcal{Q}_1 - E(\mathcal{Q}_1)) = 2\sum_{t=1}^{M} \langle E(X_t^+), X_t^+ - E(X_t^+) \rangle$$
$$\Delta(Q_2, \mathcal{Q}_2) := (Q_2 - E(Q_2)) - (\mathcal{Q}_2 - E(\mathcal{Q}_2)) = 2\sum_{t=1}^{M} \langle E(X_t^-), X_t^- - E(X_t^-) \rangle$$

For any $t > 0$, we have

$$\begin{aligned}
P^*\Big\{\frac{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z})}{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}^*)} > t\Big\} &= P^*\Big\{(Q_1 - EQ_1) - (Q_2 - EQ_2) > M(\log t) - E(Q_1 - Q_2)\Big\} \\
&\leq P^*\Big\{Q_1 - EQ_1 > \frac{M\log t - E(Q_1 - Q_2)}{2}\Big\} + P^*\Big\{Q_2 - EQ_2 < -\frac{M\log t - E(Q_1 - Q_2)}{2}\Big\} \\
&= P^*\Big\{\mathcal{Q}_1 - E\mathcal{Q}_1 > \frac{M\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_1, \mathcal{Q}_1)\Big\} \\
&+ P^*\Big\{\mathcal{Q}_2 - E\mathcal{Q}_2 < -\frac{M\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_2, \mathcal{Q}_2)\Big\}
\end{aligned}$$

where

$$P^*\Big\{\mathcal{Q}_1 - E\mathcal{Q}_1 > \frac{M\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_1, \mathcal{Q}_1)\Big\} \tag{2.22}$$
$$\leq \frac{1}{2}P^*\Big\{|\mathcal{Q}_1 - E\mathcal{Q}_1| > \frac{M\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_1, \mathcal{Q}_1)\Big\}$$
$$P^*\Big\{\mathcal{Q}_2 - E\mathcal{Q}_2 > \frac{M\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_2, \mathcal{Q}_2)\Big\}$$
$$\leq \frac{1}{2}P^*\Big\{|\mathcal{Q}_2 - E\mathcal{Q}_2| > \frac{M\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_2, \mathcal{Q}_2)\Big\}. \tag{2.23}$$

Next, we estimate each of the term in (2.22). Given the $\{Y_{ij}^t\}_{t=1}^M$ are binary random variables and the setting that any two within-community edges $Y_{i_1j_1}$ and $Y_{i_2j_2}$ have a nonnegative correlation $corr(Y_{i_1j_1}, Y_{i_2j_2}) \geq 0$. Notice that

$$corr\big(f_1(Y_{i_1j_1}), f_1(Y_{i_2j_2})\big) = \begin{cases} corr(Y_{i_1j_1}, Y_{i_2j_2}) \ \text{ if } \ \mu_{z_iz_j} \geq \mu_{z_i^*z_j^*} \\ -corr(Y_{i_1j_1}, Y_{i_2j_2}) \ \text{ if } \ \mu_{z_iz_j} < \mu_{z_i^*z_j^*} \end{cases}.$$

We denote the covariance matrix of X_t^+ and X_t^- as Σ_1 and Σ_2. Notice that a term in (2.20) is zero only when its corresponding node membership is misclassified. Define the the number of nonzero term in (1) as N_r given $\|z - z^*\|_0 = r$. Then we have $N_r = \frac{1}{2}rNM$. According to Lemma 1, X_t^+ is a subgaussian vector with a bounded subgaussian norm $L \leq C_1$ where C_1 is a positive constant and

$$L = inf\{\alpha \geq 0 : E(\exp(\langle z, X_t^+ - E(X_t^+)\rangle)) \leq \exp\{\alpha^2\langle\Sigma_1 z, z\rangle\}/2\}. \tag{2.24}$$

Next we estimate $\|\Sigma_1\|_F, \|\Sigma_1\|_{op}$ and $\|\Sigma_2\|_F, \|\Sigma_2\|_{op}$ where $\|\cdot\|_F$ is the matrix Frobenius norm and $\|\cdot\|_{op}$ is the matrix spectral norm. Denote

$$\Lambda = diag\big(\sqrt{\mathrm{Var}\{(X_t^+)_{12}\}}, \sqrt{\mathrm{Var}\{(X_t^+)_{13}\}}, \cdots, \sqrt{\mathrm{Var}\{(X_t^+)_{N-1,N}\}}\big).$$

Then $\|\Sigma_1\|_{op} = \|\Lambda R \Lambda\|_{op} \leq C_2 \|R\|_{op}$ where R is the correlation matrix of X_t^+ and based on (C1),

$$C_2 \leq \max_{1 \leq i < j \leq n} \mathrm{Var}\{(X_t^+)_{ij}\} \leq \eta_N \max\{\log \frac{\zeta}{1-\zeta}, \log \frac{1-\zeta}{\zeta}\}.$$

Denote the largest eigenvalue of R as λ_R. From the Gershgorin circle theorem, we have

$$\lambda_R \leq 1 + \max_{i=1,\cdots,N(N-1)/2} \sum_{j \neq i} |R_{ij}|.$$

Denote the number of node in the largest community is N_k. Note that the misclassification number of node $\|z - z^*\|_0 = r$ and edgewise correlation density λ both affect the sparsity of R, we have for each row in R:

$$\sum_{j \neq i} |R_{ij}| \leq \rho N_k \min(r, \lambda N_k) \leq \rho \kappa_2 N \min(r, \kappa_2 \lambda N),$$

where $\rho = \max_{i,j} R_{ij}$. Therefore, we have

$$\|\Sigma_1\|_{op} \leq C\{1 + \rho \kappa_2 \eta_N N \min(r, \kappa_2 \lambda N)\},$$

for some constant C. Similarly we have a same upper bound for $\|\Sigma_2\|_{op}$. Notice that the dimension of R is $N_r \times N_r$ and $N_r \leq rN$. In each row of R, the number of non-zero elements is less than $1 + N_k \min(r, \lambda N_k)$. Therefore, we have

$$\|\Sigma_1\|_F^2 \leq C_2 \rho^2 r \eta_N N\{1 + \kappa_2 \eta_N N \min(r, \kappa_2 \lambda N)\}.$$

Then we are able to estimate the upper bound for the first term in (2.22). According to the generalized Hanson-Wright inequality in ([30]), we have:

$$\frac{1}{2} P^*\Big\{|Q_1 - EQ_1| > s\Big\} \leq \exp\Big\{ - C \min\big(\frac{s^2}{L^4 \|\Sigma_1\|_F^2 \|A\|_F^2}, \frac{s}{L^2 \|\Sigma_1\|_{op} \|A\|_{op}}\big)\Big\}. \tag{2.25}$$

where $s = \frac{M\log t - E(Q_1-Q_2)}{2} - \Delta(Q_1, \mathcal{Q}_1)$, $A = \boldsymbol{I}_{M\times M}$ and L is subgaussian norm of X_t^+ defined in (2.24). Then we have $L \le C_1$ and $\|A\|_F^2 = M, \|A\|_{op} = 1$. To estimate s, notice

$$\begin{aligned} E(Q_1 - Q_2) &= E[\sum_{t=1}^{M}\sum_{i<j}\big\{Y_{ij}^t \log\frac{\mu_{z_iz_j}}{\mu_{z_i^*z_j^*}} + (1-Y_{ij}^t)\log\frac{1-\mu_{z_iz_j}}{1-\mu_{z_i^*z_j^*}}\big\}] \\ &= -M\sum_{i<j}\big\{\mu_{z_i^*z_j^*}\log\frac{\mu_{z_i^*z_j^*}}{\mu_{z_iz_j}} + (1-\mu_{z_i^*z_j^*})\log\frac{1-\mu_{z_i^*z_j^*}}{1-\mu_{z_iz_j}}\big\}, \end{aligned}$$

where there are total N_r non-zero terms in the summation. We introduce the function

$$k(x,y) = x\log(x/y) + (1-x)\log(1-x)/(1-y).$$

Notice that $k(x,y) > 0$ for every $x, y \in (0,1)$. Then we define:

$$c^* := \min\{k(c_{ql}, c_{q'l'})\} > 0 \tag{2.26}$$

where the minimum are taken over $\{((q,l),(q',l'))\,|c_{q,l} \ne c_{q',l'}\}$. Given that $\eta_N = o_N(1)$, it can be shown that $k(\mu_{ql}, \mu_{q'l'}) \asymp \eta_N k(c_{ql}, c_{q'l'})$. Combined with $N_r = \frac{1}{2}rNM$, we have $-E(Q_1 - Q_2) > \frac{c^*}{2}r\eta_N NM$. To estimate $\Delta(Q_1, \mathcal{Q}_1)$, given all the elements in X_t^+ are bounded, we denote $\omega_1 = \max_{1\le i<j\le n} E\{(X_t^+)_{ij}\}, \omega_2 = \max_{1\le i<j\le n}\mathrm{Var}\{(X_t^+)_{ij}\}$

$$\begin{aligned} P\big(|\Delta(Q_1,\mathcal{Q}_1)| > \frac{c^*}{2}rNM\big) &\le P\big(\omega_1|\sum_{t=1}^{M}\sum_{i=1}^{N_r}(X_{ti}^+ - E(X_{ti}^+)| > \frac{c^*}{2}rNM\big) \le \frac{\omega_1^2 M\mathrm{Var}(\sum_{i=1}^{N_r}X_{ti}^+)}{c^{*2}r^2N^2M^2/4} \\ &\le \frac{\omega_1^2(\omega_2 rN(2+\rho\lambda rN))}{c^{*2}r^2N^2M} \le O(\frac{\eta_N}{M}) \end{aligned}$$

Therefore, as M or N increases s is dominated by $-E(Q_1 - Q_2)$ with probability approaching 1. Then for any fixed $t > 0$, $s > \mathcal{O}_N(\frac{c^*}{2}rNM)$. Therefore, we have

$$\begin{aligned} &\min\big(\frac{s^2}{L^4\|\Sigma_1\|_F^2\|A\|_F^2}, \frac{s}{L^2\|\Sigma_1\|_{op}\|A\|_{op}}\big) \\ \ge &\min\big(\frac{(\frac{c^*}{2}r\eta_N NM)^2}{C_1^2MC_2\rho^2rN\{1+\kappa_2\eta_N N\min(r,\kappa_2\lambda N)\}}, \frac{\frac{c^*}{2}r\eta_N NM}{C_1C_2\{1+\rho\kappa_2\eta_N N\min(r,\kappa_2\lambda N)}\big) \\ \ge &C_3\frac{c^*r\eta_N NM}{1+\rho\kappa_2\eta_N N\min(r,\kappa_2\lambda N)}. \end{aligned}$$

where $C_3 = \frac{c^*}{C_1^2 C_2 \rho^2}$. Hence for (2.25) we have:

$$\frac{1}{2}P^*\Big\{|Q_1 - EQ_1| > s\Big\} \leq \exp\Big\{ - C\frac{c^* r\eta_N NM}{1+\rho\kappa_2\eta_N N\min(r,\kappa_2\lambda N)}\Big\},$$

where C is a positive constant. Follow Lemma 1, X_t^- is also subgaussian vector. Then we can obtain a same upper bound for

$$\frac{1}{2}P^*\Big\{|Q_2 - EQ_2| > \frac{M\log t - E(Q_1 - Q_2)}{2}\Big\}$$

in (2.22) through the above procedure. Therefore,

$$P^*\Big\{\frac{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z})}{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z^*})} > t\Big\} \leq \exp\Big\{ - C\frac{c^* r\eta_N NM}{1+\rho\kappa_2\eta_N N\min(r,\kappa_2\lambda N)}\Big\}.$$

2.10.3 Proof of Corollary 2.1

Given Theorem 2.1, we have

$$\begin{aligned}
&P_{Z*}\Big\{\sup_{\{z\neq z^*\}}\frac{L_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=z;\Theta)}{L_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=z^*;\Theta)} > t\Big\} \leq P_{Z*}\Big\{\sum_{r=1}^{N}\sum_{\|z-z^*\|_1=r}\frac{L_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=z;\Theta)}{L_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=z^*;\Theta)} > t\Big\}\\
\leq&\sum_{r=1}^{N}P_{Z*}\Big\{\sum_{\|z-z^*\|_1=r}\frac{L_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=z;\Theta)}{L_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=z^*;\Theta)} > t\Big\}\\
\leq&\sum_{r=1}^{N}\binom{N}{r}(K-1)^r\exp\Big\{ - C\frac{c^* r\eta_N NM}{1+\rho\kappa_2\eta_N N\min(r,\kappa_2\lambda N)}\Big\}\\
\leq&\sum_{r=1}^{N}\binom{N}{r}\Big\{(K-1)\exp\Big\{ - C\frac{c^*\eta_N NM}{1+\lambda\eta_N N^2}\Big\}\Big\}^r \leq (1+\Big\{(K-1)\exp\Big\{ - C\frac{c^*\eta_N NM}{1+\lambda\eta_N N^2}\Big\}\Big\})^N - 1\\
\asymp&\mathcal{O}N\exp\Big\{ - C\frac{c^*\eta_N NM}{1+\lambda\eta_N N^2}\Big\}
\end{aligned}$$

2.10.4 Proof of Theorem 2.2

We continue use the notations in the previous proof of Theorem 2.1. First decompose the proposed approximate likelihood in two parts:

$$\log\frac{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z})}{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}^*)}=\log\frac{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z})}{P_{ind}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}^*)}$$
$$+\frac{1}{M}\sum_{t=1}^{M}\log\frac{1+\sum_{k=1}^{K}\max\Big\{\sum\limits_{\substack{i<j;u<v\\(i,j)\neq(u,v)}}^{N}z_{ik}z_{jk}z_{uk}z_{vk}\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k},0\Big\}}{1+\sum_{k=1}^{K}\max\Big\{\sum\limits_{\substack{i<j;u<v\\(i,j)\neq(u,v)}}^{N}z_{ik}^*z_{jk}^*z_{uk}^*z_{vk}^*\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k},0\Big\}}.$$

Notice that ρ_{ijuv} is the empirical estimator based in $\hat{Y}_{ij}^{t,k}$ and $\hat{Y}_{uv}^{t,k}$, then $\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}>0$ with high probability. Based on the mean value theorem, we have for some constant C_1 that

$$\log\frac{1+\sum_{k=1}^{K}\max\Big\{\sum\limits_{\substack{i<j;u<v\\(i,j)\neq(u,v)}}^{N}z_{ik}z_{jk}z_{uk}z_{vk}\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k},0\Big\}}{1+\sum_{k=1}^{K}\max\Big\{\sum\limits_{\substack{i<j;u<v\\(i,j)\neq(u,v)}}^{N}z_{ik}^*z_{jk}^*z_{uk}^*z_{vk}^*\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k},0\Big\}}$$
$$=C_1\sum_{k=1}^{K}\Big\{\max\Big(\sum_{\substack{i<j;u<v\\(i,j)\neq(u,v)}}^{N}z_{ik}z_{jk}z_{uk}z_{vk}\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k},0\Big)-\max\Big(\sum_{\substack{i<j;u<v\\(i,j)\neq(u,v)}}^{N}z_{ik}^*z_{jk}^*z_{uk}^*z_{vk}^*\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k},0\Big)\Big\}$$
$$\leq C_1\sum_{k=1}^{K}\Big\{\sum_{\substack{i<j;u<v\\(i,j)\neq(u,v)}}^{N}(z_{ik}z_{jk}z_{uk}z_{vk}-z_{ik}^*z_{jk}^*z_{uk}^*z_{vk}^*)\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}\Big\}. \tag{2.27}$$

Notice in summation (2.27), the terms are non-zero only when $z_{ik}z_{jk}z_{uk}z_{vk}\neq z_{ik}^*z_{jk}^*z_{uk}^*z_{vk}^*$. We denote two node sets

$$\xi_1=\{(i,j,u,v)|z_{ik}z_{jk}z_{uk}z_{vk}=1,z_{ik}^*z_{jk}^*z_{uk}^*z_{vk}^*=0,k=1,\cdots,K\},$$
$$\xi_2=\{(i,j,u,v)|z_{ik}^*z_{jk}^*z_{uk}^*z_{vk}^*=1,z_{ik}z_{jk}z_{uk}z_{vk}=0,k=1,\cdots,K\}.$$

where $\#|\xi_1| = N_1$ and $\#|\xi_2| = N_2$. Given the number of misclassified nodes $\|z - z^*\|_0 = r$, we have $N_1 = \mathcal{O}(rN^3)$ and $N_2 = \mathcal{O}(rN^3)$. In the following, we construct the augmented edge vectors for the t th sample network by incorporating the vectorized pairwise edge interaction in (2.27) such that:

$$\tilde{X}_t^+ = \Big\{ X_t^+, \underbrace{\Big(\sqrt{\frac{C_1}{2}\{\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}\}_+}\Big)_{1\times N_1}}_{\substack{(i,j,u,v)\in\xi_1\\ z_{ik}z_{jk}z_{uk}z_{vk}=1\\ k=1,\cdots,K}}, \underbrace{\Big(\sqrt{\frac{C_1}{2}\{-\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}\}_+}\Big)_{1\times N_2}}_{\substack{(i,j,u,v)\in\xi_2\\ z^*_{ik}z^*_{jk}z^*_{uk}z^*_{vk}=1\\ k=1,\cdot,K}} \Big\},$$

$$\tilde{X}_t^- = \Big\{ X_t^-, \underbrace{\Big(\sqrt{\frac{C_1}{2}\{\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}\}_-}\Big)_{1\times N_1}}_{\substack{(i,j,u,v)\in\xi_1\\ z_{ik}z_{jk}z_{uk}z_{vk}=1\\ k=1,\cdots,K}}, \underbrace{\Big(\sqrt{\frac{C_1}{2}\{-\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}\}_-}\Big)_{1\times N_2}}_{\substack{(i,j,u,v)\in\xi_2\\ z^*_{ik}z^*_{jk}z^*_{uk}z^*_{vk}=1\\ k=1,\cdots,K}} \Big\}.$$

where X_t^+ and X_t^- are defined in (2.21). Denote the covariance matrix for $\tilde{X}_t^+$ and $\tilde{X}_t^-$ are $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$ respectively. Since the second-order terms in X_t^+ and X_t^- such as $\sqrt{\frac{C_1}{2}\{\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}\}_+}$ only rescale the original edgewise interaction $\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}$ then they preserve the third-order and fourth-order correlation within communities such that

$$|E\Big\{ f_1(Y_{i_1j_1}^t)\sqrt{\frac{C}{2}\{\rho_{ijuv}\hat{Y}_{i_2j_2}^{t,k}\hat{Y}_{i_3j_3}^{t,k}\}_+}\Big\}| = |E(\hat{Y}_{i_1j_1}^{t,k}\hat{Y}_{i_2j_2}^{t,k}\hat{Y}_{i_3j_3}^{t,k})|,$$

$$|E\Big\{ f_2(Y_{i_1j_1}^t)\sqrt{\frac{C}{2}\{\rho_{ijuv}\hat{Y}_{i_2j_2}^{t,k}\hat{Y}_{i_3j_3}^{t,k}\}_-}\Big\}| = |E(\hat{Y}_{i_1j_1}^{t,k}\hat{Y}_{i_2j_2}^{t,k}\hat{Y}_{i_3j_3}^{t,k})|,$$

$$|E\Big\{ \sqrt{\frac{C}{2}\{\rho_{ijuv}\hat{Y}_{i_1j_1}^{t,k}\hat{Y}_{i_2j_2}^{t,k}\}_+}\sqrt{\frac{C}{2}\{\rho_{ijuv}\hat{Y}_{i_3j_3}^{t,k}\hat{Y}_{i_4j_4}^{t,k}\}_+}\Big\}| = |E(\hat{Y}_{i_1j_1}^{t,k}\hat{Y}_{i_2j_2}^{t,k}\hat{Y}_{i_3j_3}^{t,k}\hat{Y}_{i_4j_4}^{t,k})|,$$

$$|E\Big\{ \sqrt{\frac{C}{2}\{\rho_{ijuv}\hat{Y}_{i_1j_1}^{t,k}\hat{Y}_{i_2j_2}^{t,k}\}_-}\sqrt{\frac{C}{2}\{\rho_{ijuv}\hat{Y}_{i_3j_3}^{t,k}\hat{Y}_{i_4j_4}^{t,k}\}_-}\Big\}| = |E(\hat{Y}_{i_1j_1}^{t,k}\hat{Y}_{i_2j_2}^{t,k}\hat{Y}_{i_3j_3}^{t,k}\hat{Y}_{i_4j_4}^{t,k})|.$$

Notice that each element in $\tilde{X}_t^+$ or $\tilde{X}_t^-$ is a bounded binary random variable. Follow the same procedure in Lemma 1, we can show that both $\tilde{X}_t^+$ and $\tilde{X}_t^-$ are subgaussian random vectors such that $L_1 \le C_2, L_2 \le C_2$ for some constant C_2 where L_1, L_2 are subgaussian norm of $\tilde{X}_t^+$ and $\tilde{X}_t^-$.

Then consider the following quadratic forms:

$$\tilde{Q}_1 = \sum_{t=1}^{M} \langle \tilde{X}_t^+, \tilde{X}_t^+ \rangle, \tilde{Q}_2 = \sum_{t=1}^{M} \langle \tilde{X}_t^-, \tilde{X}_t^- \rangle.$$

Therefore, we have

$$\log \frac{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z})}{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}^*)} \leq \frac{1}{M}(\tilde{Q}_1 - \tilde{Q}_2).$$

Denote the centralized version quadratic forms $\tilde{Q}_1$ and $\tilde{Q}_2$ as $\tilde{\mathcal{Q}}_1$ and $\tilde{\mathcal{Q}}_2$ such that

$$\tilde{\mathcal{Q}}_1 = \sum_{t=1}^{M} \langle \tilde{X}_t^+ - E(\tilde{X}_t^+), \tilde{X}_t^+ - E(\tilde{X}_t^+) \rangle, \tilde{\mathcal{Q}}_2 = \sum_{t=1}^{M} \langle \tilde{X}_t^- - E(\tilde{X}_t^-), \tilde{X}_t^- - E(\tilde{X}_t^-) \rangle.$$

Denote the following quadratic difference as:

$$\Delta(\tilde{Q}_1, \tilde{\mathcal{Q}}_1) := (\tilde{Q}_1 - E(\tilde{Q}_1)) - (\tilde{\mathcal{Q}}_1 - E(\tilde{\mathcal{Q}}_1)) = 2\sum_{t=1}^{M} \langle E(\tilde{X}_t^+), \tilde{X}_t^+ - E(\tilde{X}_t^+) \rangle$$
$$\Delta(\tilde{Q}_2, \tilde{\mathcal{Q}}_2) := (\tilde{Q}_2 - E(\tilde{Q}_2)) - (\tilde{\mathcal{Q}}_2 - E(\tilde{\mathcal{Q}}_2)) = 2\sum_{t=1}^{M} \langle E(\tilde{X}_t^-), \tilde{X}_t^- - E(\tilde{X}_t^-) \rangle$$

Similar to (2.22), for any fixed $t > 0$:

$$\begin{aligned}
&P^*\Big\{ \frac{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z})}{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z}^*)} > t \Big\} \leq P^*\Big\{ \frac{1}{M}(\tilde{Q}_1 - \tilde{Q}_2) > \log t \Big\} \\
\leq & P^*\Big\{ \tilde{Q}_1 - E\tilde{Q}_1 > \frac{M\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} \Big\} + P^*\Big\{ \tilde{Q}_2 - E\tilde{Q}_2 < -\frac{M\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} \Big\} \\
= & P^*\Big\{ \tilde{\mathcal{Q}}_1 - E\tilde{\mathcal{Q}}_1 > \frac{M\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} - \Delta(\tilde{Q}_1, \tilde{\mathcal{Q}}_1) \Big\} \\
& + P^*\Big\{ \tilde{\mathcal{Q}}_2 - E\tilde{\mathcal{Q}}_2 < -\frac{M\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} - \Delta(\tilde{Q}_2, \tilde{\mathcal{Q}}_2) \Big\} \\
\leq & \frac{1}{2}P^*\Big\{ |\tilde{\mathcal{Q}}_1 - E\tilde{\mathcal{Q}}_1| > \frac{M\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} - \Delta(\tilde{Q}_1, \tilde{\mathcal{Q}}_1) \Big\} \\
& + \frac{1}{2}P^*\Big\{ |\tilde{\mathcal{Q}}_2 - E\tilde{\mathcal{Q}}_2| > \frac{M\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} - \Delta(\tilde{Q}_2, \tilde{\mathcal{Q}}_2) \Big\}. \qquad (2.28)
\end{aligned}$$

Next we estimate $\|\tilde{\Sigma}_1\|_F$, $\|\tilde{\Sigma_1}\|_{op}$ and $\|\tilde{\Sigma}_2\|_F$, $\|\tilde{\Sigma}_2\|_{op}$. Denote

$$\tilde{\Lambda} = diag(\Lambda, \underbrace{sd\Big(\sqrt{\frac{1}{2}\{\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}\}_+}\Big)_{1\times N_1}}_{\substack{(i,j,u,v)\in\xi_1\\ z_{ik}z_{jk}z_{uk}z_{vk}=1\\ k=1,\cdots,K}}, \underbrace{sd\Big(\sqrt{\frac{1}{2}\{-\rho_{ijuv}\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}\}_+}\Big)_{1\times N_2}}_{\substack{(i,j,u,v)\in\xi_2\\ z^*_{ik}z^*_{jk}z^*_{uk}z^*_{vk}=1\\ k=1,\cdots,K}}),$$

then $\|\tilde{\Sigma}_1\|_{op} = \|\tilde{\Lambda}\tilde{R}\tilde{\Lambda}\|_{op} \leq C_3\|\tilde{R}\|_{op}$ where $\tilde{R}$ is the correlation matrix of $\tilde{X}_t^+$ and C_3 is the largest variance of elements in $\tilde{X}_t^+$. Denote the largest eigenvalue of $\tilde{R}$ as $\lambda_{\tilde{R}}$. From the Gershgorin circle theorem, we have

$$\lambda_{\tilde{R}} \leq 1 + \max_i \sum_{j\neq i} \left|\tilde{R}_{ij}\right|.$$

Given that the misclassification number of node $\|z - z^*\|_0 = r$, edgewise correlation density λ and condition C3, for each row in $\tilde{R}$, there exists some constant C_4>0 such that:

$$\sum_{j\neq i} |R_{ij}| \leq C_4\rho N_k \min(r, \lambda N_k) = C_4\rho\kappa_2 N \min(r, \kappa_2\lambda N), \tag{2.29}$$

where $\rho = \max_{i,j} \tilde{R}_{ij}$. Therefore, we have

$$\|\tilde{\Sigma}_1\|_{op} \leq C_3\{1 + C_4\rho\kappa_2 N \min(r, \kappa_2\lambda N)\}.$$

Similarly, $\|\tilde{\Sigma}_2\|_{op}$ follows a same upper bound. Notice that the dimension of $\tilde{R}$ is $(N_r + N_1 + N_2) \times (N_r + N_1 + N_2)$. Under the condition C3, in each row of $\tilde{R}$, the number of non-zero elements is less than $1 + C_4 N_k \min(r, \lambda N_k)$. Therefore, we have for a constant $C' > 0$:

$$\begin{aligned}\|\tilde{\Sigma}_1\|_F^2 \leq& C_3\rho^2(N_r + N_1 + N_2)\{1 + C_4\kappa_2 N \min(r, \kappa_2\lambda N)\}\\ \leq& C'\rho^2(rN + rN^3)\{1 + C_4\kappa_2 N \min(r, \kappa_2\lambda N)\}.\end{aligned}$$

According to the generalized Hanson-Wright inequality in ([30]):

$$\frac{1}{2}P^*\Big\{|\tilde{Q}_1 - E\tilde{Q}_1| > s\Big\} \leq \exp\Big\{-C\min\big(\frac{s^2}{L_1^4\|\tilde{\Sigma}_1\|_F^2\|A\|_F^2}, \frac{s}{L_1^2\|\tilde{\Sigma}_1\|_{op}\|A\|_{op}}\big)\Big\}, \tag{2.30}$$

where $s = \frac{M\log t - E(\tilde{\mathcal{Q}}_1 - \tilde{\mathcal{Q}}_2)}{2} - \Delta(\tilde{Q}_1, \tilde{\mathcal{Q}}_1)$, $A = \boldsymbol{I}_{M\times M}$ and L_1 is subgaussian norm of $\tilde{X}_t^+$. Notice $\|A\|_F^2 = M, \|A\|_{op} = 1$. Given (2.27), we have

$$E(\tilde{Q}_1 - \tilde{Q}_2) = E(Q_1 - Q_2) + C_1 \sum_{k=1}^{K} \Big\{ \sum_{\substack{i<j;u<v \\ (i,j)\neq(u,v)}}^{N} (z_{ik}z_{jk}z_{uk}z_{vk} - z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^*)\rho_{ijuv} E(\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}) \Big\}.$$

Denote ρ_{min} as the lower bound of all non-zero correlation among edges such that $E(\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}) = \rho_{ijuv} \geq \rho_{min}$. Given the edges from different communities are independent and within-community correlation density λ, we have for some positive constant C_5,

$$\#|\{(i,j,u,v) : E(\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}) \neq 0, (i,j,u,v) \in \xi_2\}| = \lambda N_1 = \lambda C_5 r N^3,$$

$$\#|\{(i,j,u,v) : E(\hat{Y}_{ij}^{t,k}\hat{Y}_{uv}^{t,k}) \neq 0, (i,j,u,v) \in \xi_1\}| \leq \lambda \binom{r}{4}.$$

Assume that $r \leq cN$ for some constant $0 < c < 1$, we have for some constant $c_0 > 0$:

$$-E(\tilde{Q}_1 - \tilde{Q}_2) \geq \frac{c^*}{2} rNM + \lambda M \frac{C_1 \rho_{min}^2}{2} (C_5 r N^3 - \binom{r}{4}) \geq c_0 r (c^* \eta_N N + \lambda N^3) M.$$

To estimate $\Delta(\tilde{Q}_1, \tilde{\mathcal{Q}}_1)$, given all the elements in $\tilde{X}_t^+$ are bounded, we denote

$$\omega_3 = \max_i E\{(\tilde{X}_t^+)_i\}, \omega_4 = \max_i \mathrm{Var}\{(\tilde{X}_t^+)_i\}.$$

According to the definition of $\tilde{X}_t^+$ and $N_1 = \mathcal{O}(rN^3), N_2 = \mathcal{O}(rN^3)$, there exists a positive constant C^+ such that $\#|\tilde{X}_t^+| = \frac{rN}{2} + C^+ r N^3$, therefore

$$\begin{aligned} & P\big(|\Delta(\tilde{Q}_1, \tilde{\mathcal{Q}}_1)| > c_0 r (c^* \eta_N N + \lambda N^3) M\big) \\ & \leq P\big(\omega_3 | \sum_{t=1}^{M} \sum_{i=1}^{\#|\tilde{X}_t^+|} (\tilde{X}_{ti}^+ - E(\tilde{X}_{ti}^+)| > c_0 r (c^* \eta_N N + \lambda N^3) M\big) \\ & \leq \frac{\omega_3^2 M \mathrm{Var}(\sum_{i=1}^{\#|\tilde{X}_t^+|} \tilde{X}_{ti}^+)}{c_0^2 r^2 (c^* \eta_N N + \lambda N^3)^2 M^2} \end{aligned} \tag{2.31}$$

From the assumption (C3), there exists a positive constant ω_5 such that

$$\begin{aligned}\mathrm{Var}(\sum_{i=1}^{\#|\tilde{X}_t^+|}\tilde{X}_{ti}^+) &= \sum_{i=1}^{\#|\tilde{X}_t^+|}\mathrm{Var}(\tilde{X}_{ti}^+) + \sum_{i,j}\mathrm{Cov}(\tilde{X}_{ti}^+, \tilde{X}_{tj}^+) \\ \leq &\omega_4(\frac{rN}{2} + C^+rN^3) + w_4\rho(\frac{\lambda r^2N^2}{4} + rN^3\cdot\omega_5\lambda N^2 + rN\cdot\omega_5\lambda N^2)\end{aligned} \tag{2.32}$$

Through combining (2.31) and (2.32), give $\lambda > 0$ we have

$$P\big(|\Delta(\tilde{Q}_1, \tilde{\mathcal{Q}}_1)| > c_0r(c^*\eta_N N + \lambda N^3)M\big) \leq \frac{\omega_3^2\omega_4}{c_0^2M}(\frac{1}{2r\lambda^2N^5} + \frac{C^+}{r\lambda^2N^3} + \frac{\rho}{4\lambda N^4} + \frac{\rho\omega_5}{r\lambda N} + \frac{\rho\omega_5}{r\lambda N^3})$$

Therefore, given $N > \mathcal{O}_N(\frac{1}{\lambda})$ and M, N increasing, s is dominated by $-E(\tilde{Q}_1 - \tilde{Q}_2)$ with probability approaching 1. Given any fixed $t > 0$, $s > \mathcal{O}_N(r(c^*\eta_N N + \lambda N^3)M)$. For the first term in (2.30),

$$\frac{s^2}{L_1^4\|\tilde{\Sigma}_1\|_F^2\|A\|_F^2} \geq \frac{r^2(c^*\eta_N N + \lambda N^3)^2M^2}{L_1^4C'\rho^2(rN + rN^3)\{1 + C_4\kappa_2N\min(r, \kappa_2\lambda N)\}M}.$$

For the second term in (2.30),

$$\frac{s}{L_1^2\|\tilde{\Sigma}_1\|_{op}\|A\|_{op}} \geq \frac{r(c^*\eta_N N + \lambda N^3)M}{L_1^2C'\{1 + C_4\rho\kappa_2N\min(r, \kappa_2\lambda N)\}}.$$

Given $\lambda > 0$, we have for some constant $C_6 > 0$

$$\min\big(\frac{s^2}{L_1^4\|\tilde{\Sigma}_1\|_F^2\|A\|_F^2}, \frac{s}{L_1^2\|\tilde{\Sigma}_1\|_{op}\|A\|_{op}}\big) \geq C_6\frac{r\lambda NM(c^*\eta_N + \lambda N^2)}{1 + C_4\rho\kappa_2N\min(r, \kappa_2\lambda N)}. \tag{2.33}$$

Follow the same procedure we can show a upper bound for $P^*\Big\{|\tilde{Q}_2 - E\tilde{Q}_2| > s\Big\}$ with a same order to (2.33). Combined with (2.28) and (2.30), we have for $\lambda > 0$ and some constant $C > 0$:

$$P_{Z^*}\Big\{\frac{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z} = z; \Theta)}{\tilde{L}(\boldsymbol{Y}|\boldsymbol{Z} = z^*; \Theta)} > t\Big\} \leq \exp\Big\{-C\frac{r\lambda NM(c^*\eta_N + \lambda N^2)}{1 + C_4\rho\kappa_2N\min(r, \kappa_2\lambda N)}\Big\},$$

2.10.5 Proof of Corollary 2.2

The proof follows a similar discussion for Corollary 2.1.

2.10.6 Proof of Theorem 2.3

Follow the notations introduced in Theorem 2.1 and Theorem 2.2, we further define that $\boldsymbol{w} = \max \frac{P^{(s)}(Z_i=q)}{P^{(s)}(Z_i=l)},\ i = 1, \cdots, N,\ q, l = 1, \cdots, K$. Let $\boldsymbol{E}$ stands for the operator of expectation step in Algorithm 1 in Section 4.

We first consider the misclassification of updated estimated membership for node s, e.g., $\boldsymbol{E}(z_s)$ from the current estimation α_s. We denote that $\boldsymbol{\alpha}_{-s}$ as the probability estimations of nodes' memberships at current step except node s and assume the true membership of node s is b, i.e., $z_s^* = b$. If we use the marginal likelihood, then:

$$
\begin{aligned}
&\|\boldsymbol{E}(z_s) - z_s^*\|_1 = \\
&|\frac{P(z_s=1)\tilde{L}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s=1)}{\sum_{q=1}^K P(z_s=q)\tilde{L}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s=q)} - 0| + \cdots + |\frac{P(z_s=b)\tilde{L}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s=b)}{\sum_{q=1}^K P(z_s=K)\tilde{L}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s=K)} - 1| \\
&\leq 2\frac{\sum_{q\neq b} P(z_s=q)\tilde{L}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s=q)}{\sum_{q=1}^K P(z_s=q)\tilde{L}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; Z_s=q)} \leq 2\boldsymbol{w}\sum_{q\neq b}^K \frac{\tilde{L}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s=q)}{\tilde{L}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s=b)} \\
&= 2\boldsymbol{w}\sum_{q\neq b}^K \min[1, \exp\{\log\tilde{L}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; Z_s=q) - \log\tilde{L}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s=b)\}]. \qquad (2.34)
\end{aligned}
$$

Then given node s belongs to different communities while the estimated membership for other nodes $\boldsymbol{\alpha}_{-s}$ are fixed. We decompose the proposed approximate likelihood into marginal part and correlation part in the following: $\log\tilde{L}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s) = \log L_{mar}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s) + \log L_{cor}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s)$. The marginal likelihood $\log L_{mar}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s)$,

$$
\begin{aligned}
&\log L_{mar}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s = a) \\
&= \frac{1}{M}\sum_{t=1}^M \Bigg[\underbrace{log\prod_{q,l}^K \prod_{i\neq j\neq s}^N \left\{\mu_{ql}^{Y_{ij}^t}(1-\mu_{ql})^{(1-Y_{ij}^t)}\right\}^{\alpha_{iq}\alpha_{jl}}}_{\text{not depend on } z_s} + \prod_{q=1}^K \prod_{i\neq s}^N \left\{\mu_{qa}^{Y_{is}^t}(1-\mu_{qa})^{(1-Y_{is}^t)}\right\}^{\alpha_{iq}} \Bigg].
\end{aligned}
$$

Therefore, the discrepancy among marginal likelihood is

$$
\begin{aligned}
&\log L_{mar}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s = a) - \log L_{mar}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s = b) \\
=&\frac{1}{M}\sum_{t=1}^{M}\sum_{q=1}^{K}\sum_{i\neq s}^{N}\Big[\alpha_{iq}\{Y_{is}^t \log\frac{\hat{\mu}_{qa}}{\hat{\mu}_{qb}} + (1-Y_{is}^t)\log\frac{1-\hat{\mu}_{qa}}{1-\hat{\mu}_{qb}}\}\Big] \\
=&\frac{1}{M}\sum_{t=1}^{M}\sum_{q=1}^{K}\sum_{i\neq s}^{N}\Big[\alpha_{iq}\{Y_{is}^t \log\frac{\mu_{qa}}{\mu_{qb}} + (1-Y_{is}^t)\log\frac{1-\mu_{qa}}{1-\mu_{qb}}\}\Big] \\
+&\frac{1}{M}\sum_{t=1}^{M}\sum_{q=1}^{K}\sum_{i\neq s}^{N}\Big[\alpha_{iq}\{Y_{is}^t \log\frac{\mu_{qa}\hat{\mu}_{qb}}{\hat{\mu}_{qa}\mu_{qb}} + (1-Y_{is}^t)\log\frac{(1-\mu_{qa})(1-\hat{\mu}_{qb})}{(1-\hat{\mu}_{qa})(1-\hat{\mu}_{qb})}\}\Big]
\end{aligned}
$$

We can decompose the marginal discrepancy into four parts:

$$
\begin{aligned}
&\log L_{mar}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s = a) - \log L_{mar}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s = b) \\
=&\underbrace{\frac{1}{M}\sum_{t=1}^{M}\sum_{q=1}^{K}\sum_{i\neq s}^{N}(\alpha_{iq} - z_{iq}^*)\{Y_{is}^t - E(Y_{is}^t)\}(\log\frac{\mu_{qa}}{\mu_{qb}} - \log\frac{1-\mu_{qa}}{1-\mu_{qb}})}_{\boldsymbol{A_1}} \\
+&\underbrace{\frac{1}{M}\sum_{t=1}^{M}\sum_{q=1}^{K}\sum_{i\neq s}^{N}\Big[(\alpha_{iq} - z_{iq}^*)\{EY_{is}^t \log\frac{\mu_{qa}}{\mu_{qb}} + (1-EY_{is}^t)\log\frac{1-\mu_{qa}}{1-\mu_{qb}}\}\Big]}_{\boldsymbol{A_2}} \\
+&\underbrace{\frac{1}{M}\sum_{t=1}^{M}\sum_{q=1}^{K}\sum_{i\neq s}^{N}\Big[z_{iq}^*\{Y_{is}^t \log\frac{\mu_{qa}}{\mu_{qb}} + (1-Y_{is}^t)\log\frac{1-\mu_{qa}}{1-\mu_{qb}}\}\Big]}_{\boldsymbol{A_3}} \\
+&\underbrace{\frac{1}{M}\sum_{t=1}^{M}\sum_{q=1}^{K}\sum_{i\neq s}^{N}\Big[\alpha_{iq}\{Y_{is}^t \log\frac{\mu_{qa}\hat{\mu}_{qb}}{\hat{\mu}_{qa}\mu_{qb}} + (1-Y_{is}^t)\log\frac{(1-\mu_{qa})(1-\hat{\mu}_{qb})}{(1-\hat{\mu}_{qa})(1-\hat{\mu}_{qb})}\}\Big]}_{\boldsymbol{A_4}}.
\end{aligned}
$$

For the correlation part, we consider the pairwise interaction terms in the $\log L_{cor}(\boldsymbol{Y}|\boldsymbol{\alpha})$. Notice that for $t = 1, \cdots, M$

$$
\sum_{\substack{i<j;k<g \\ (i,j)\neq(k,g)}}^{N} \alpha_{ia}\alpha_{ja}\alpha_{ka}\alpha_{ga}\rho_{ijkg}\hat{Y}_{ij}^{t,a}\hat{Y}_{kg}^{t,a} = (\sum_{i\neq s}^{N}\alpha_{sa}\alpha_{ia}\hat{Y}_{si}^{t,a})(\sum_{i<j}^{N}\alpha_{ia}\alpha_{ja}\hat{Y}_{ij}^{t,a}) - \sum_{i\neq s}^{N}(\alpha_{ia}\hat{Y}_{si}^{t,a})^2 + A_a^t,
$$

where A_q^t does not depend on z_s. Since the first term $(\sum_{i\neq s}^N \alpha_{sa}\alpha_{ia}\hat{Y}_{si}^{t,a})(\sum_{i<j}^N \alpha_{ia}\alpha_{ja}\hat{Y}_{ij}^{t,a}) = o(N^3)$ and the second term $\sum_{i\neq s}^N(\alpha_{ia}\hat{Y}_{si}^{t,a})^2 = o(N)$, without loss of generality, we can keep the first dominating term when N is large. For the correlation part $\log L_{cor}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; z_s)$, if $\alpha_{sq} = 0,\ q \neq a$ and $\alpha_{sa} = 1$:

$$\begin{aligned}\log L_{cor}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; Z_s = a) =& \frac{1}{M}\sum_{t=1}^{M}\Big\{1 + \sum_{q=1}^{K}\frac{\rho_q}{2}\max(\sum_{\substack{i<j;k<g\\(i,j)\neq(k,g)}}^{N}\alpha_{iq}\alpha_{jq}\alpha_{kq}\alpha_{gq}\hat{Y}_{ij}^{t,q}\hat{Y}_{kg}^{t,q}, 0)\Big\}\\ =& 1 + \underbrace{\frac{1}{M}\sum_{t=1}^{M}\sum_{q=1}^{K}\frac{\rho_q}{2}A_q^t}_{\boldsymbol{A}} + \underbrace{\frac{\rho_a}{2}(\sum_{i\neq s}^{N}\alpha_{sa}\alpha_{ia}\hat{Y}_{si}^{t,a})(\sum_{i<j}^{N}\alpha_{ia}\alpha_{ja}\hat{Y}_{ij}^{t,a})}_{\boldsymbol{B}_a}.\end{aligned}$$

Through the Taylor expansion, the discrepancy of correlation information when node s belongs to different communities a and b:

$$\begin{aligned}\log L_{cor}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; Z_s = a) - \log L_{cor}(\boldsymbol{Y}|\boldsymbol{\alpha}_{-s}; Z_s = b) &= \log(1 + \boldsymbol{A} + \boldsymbol{B_a}) - \log(1 + \boldsymbol{A} + \boldsymbol{B_b})\\ &= \log(1 + \frac{\boldsymbol{B_a} - \boldsymbol{B_b}}{1 + \boldsymbol{A} + \boldsymbol{B_b}}) \leq \boldsymbol{C}_A(\boldsymbol{B_a} - \boldsymbol{B_b}),\end{aligned}$$

where $\boldsymbol{C}_A$ is a constant relating to the gradient of function $\log(1 + 1/x)$ at $\boldsymbol{A}$. Then we set $\rho = \min \rho_q, q = 1, \cdots, K$

$$\begin{aligned}\boldsymbol{B_a} - \boldsymbol{B_b} &= (\sum_{i\neq s}^{N}\alpha_{ia}\hat{Y}_{si}^{t,a})(\sum_{i<j}^{N}\alpha_{ia}\alpha_{ja}\hat{Y}_{ij}^{t,a}) - (\sum_{i\neq s}^{N}\alpha_{ib}\hat{Y}_{si}^{t,b})(\sum_{i<j}^{N}\alpha_{ib}\alpha_{jb}\hat{Y}_{ij}^{t,b})\\ &\leq \frac{\rho}{4}\Big(\langle\alpha_a \otimes vec(\alpha_a^T\alpha_a), \hat{Y}_{.s}^{t,a} \otimes vec(\hat{\boldsymbol{Y}}^{t,a})\rangle - \langle\alpha_b \otimes vec(\alpha_b^T\alpha_b), \hat{Y}_{.s}^{t,b} \otimes vec(\hat{\boldsymbol{Y}}^{t,b})\rangle\Big).\end{aligned}$$

For the simplicity of notation, we define and decompose the correlation discrepancy as followings:

$$\begin{aligned}\boldsymbol{B} :=& \sum_{t=1}^{M} \frac{\rho C_A}{4M} \big(\langle \alpha_a \otimes vec(\alpha_a^T \alpha_a), \hat{Y}_{\cdot s}^{t,a} \otimes vec(\hat{\boldsymbol{Y}}^{t,a}) \rangle - \langle \alpha_b \otimes vec(\alpha_b^T \alpha_b), \hat{Y}_{\cdot i}^{t,b} \otimes vec(\hat{\boldsymbol{Y}}^{t,b}) \rangle \big) \\ =& \underbrace{\frac{\rho C_A}{4M} \sum_{t=1}^{M} \big(\langle \alpha_a \otimes vec(\alpha_a^T \alpha_a) - z_a^* \otimes vec(z_a^{*T} z_a^*), \hat{Y}_{\cdot s}^{t,a} \otimes vec(\hat{\boldsymbol{Y}}^{t,a}) \rangle -}_{} \\ & \underbrace{\langle \alpha_b \otimes vec(\alpha_b^T \alpha_b) - z_b^* \otimes vec(z_b^{*T} z_b^*), \hat{Y}_{\cdot s}^{t,b} \otimes vec(\hat{\boldsymbol{Y}}^{t,b}) \rangle \big)}_{\text{misclassification error: } \boldsymbol{B}_1} \\ & + \underbrace{\frac{\rho C_A}{4M} \sum_{t=1}^{M} \big(\langle z_a^* \otimes vec(z_a^{*T} z_a^*), \hat{Y}_{\cdot s}^{t,a} \otimes vec(\hat{\boldsymbol{Y}}^{t,a}) \rangle - \langle z_b^* \otimes vec(z_b^{*T} z_b^*), \hat{Y}_{\cdot s}^{t,b} \otimes vec(\hat{\boldsymbol{Y}}^{t,b}) \rangle \big)}_{\text{estimation bias: } \boldsymbol{B}_2} .\end{aligned}$$

Notice that $\min\{1, \exp(x)\} \leq \exp(x_0) + \sum_{l=0}^{m-1} \frac{1-\exp(x_0)}{m} \mathbb{1}\{x \geq (1 - l/m)x_0\}$ and set $x_0 = -\alpha' MN$, where $\alpha' = \frac{\lambda(c^* \eta_N + \lambda N^2)}{1+\lambda N^2}$. Given (2.34), we have:

$$E\|\boldsymbol{\alpha}^{s+1} - \boldsymbol{z^*}\|_1 \leq 2\boldsymbol{w} NK \exp(-\alpha' MN) + 2\boldsymbol{w} \sum_{l=0}^{m-1} \sum_{a=1}^{K} \sum_{b \neq a} \sum_{i: z_i^* = b} \frac{1 - exp(\alpha' MN)}{m} E(\boldsymbol{L}_2) \tag{2.35}$$

where $E(\boldsymbol{L}) = \mathbb{P}\Big(\boldsymbol{A} + \boldsymbol{B} \geq \frac{m-l}{m} x_0 \Big)$. For some specific $t > 0$,

$$\begin{aligned}&\mathbb{P}\Big(\boldsymbol{A} + \boldsymbol{B} \geq \frac{m-l}{m} x_0 \Big) = \mathbb{P}\Big(\boldsymbol{A_1} + \boldsymbol{A_2} + \boldsymbol{A_3} + \boldsymbol{A_4} + \boldsymbol{B_1} + \boldsymbol{B_2} \geq \frac{m-l}{m} x_0 \Big) \\ \leq& \mathbb{P}\Big(\boldsymbol{A_1} + \boldsymbol{B_1} \geq t \Big) + \mathbb{P}\Big(\boldsymbol{A_3} + \boldsymbol{B_2} \geq \frac{m-l}{m} x_0 - t - \boldsymbol{A_2} - \boldsymbol{A_4} \Big).\end{aligned} \tag{2.36}$$

We then transfer $\boldsymbol{A_3} + \boldsymbol{B_2}$ into a quadratic form. For each community $q, q = 1, \cdots, K$ define the transformations:

$$f_q^+(x) = \sqrt{\Big[z_{iq}^* \{ Y_{is}^t \log \frac{\mu_{qa}}{\mu_{qb}} + (1 - Y_{is}^t) \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}} \} \Big]_+},$$

$$f_q^-(x) = \sqrt{\Big[z_{iq}^* \{ Y_{is}^t \log \frac{\mu_{qa}}{\mu_{qb}} + (1 - Y_{is}^t) \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}} \} \Big]_-},$$

$$X_t^+ = \{f_1^+(Y_{1s}^t), \cdots, f_1^+(Y_{ns}^t), f_2^+(Y_{1s}^t), \cdots, f_2^+(Y_{Ns}^t), \cdots, f_K^+(Y_{1s}^t), \cdots, f_K^+(Y_{Ns}^t)\},$$

$$X_t^- = \{f_1^-(Y_{1s}^t), \cdots, f_1^-(Y_{Ns}^t), f_2^-(Y_{1s}^t), \cdots, f_2^-(Y_{Ns}^t), \cdots, f_K^-(Y_{1s}^t), \cdots, f_K^-(Y_{Ns}^t)\}.$$

Notice that the total number of non-zero terms in X_t^+ or X_t^- is N. We define the node sets

$$\tilde{\xi}_a = \{(i_1, i_2, i_3) | z_{i_1 a}^* z_{i_2 a}^* z_{i_3 a}^* = 1\} \quad \tilde{\xi}_b = \{(i_1, i_2, i_3) | z_{i_1 b}^* z_{i_2 b}^* z_{i_3 b}^* = 1\}.$$

Note $\#|\tilde{\xi}_a| = o(N_a^3)$ and $\#|\tilde{\xi}_b| = o(N_b^3)$ where N_a and N_b are number of node in community a and b. We further define augmented edges vectors:

$$\bar{X}_t^+ = \Bigg(X_t^+, \underbrace{\Big(\frac{C_A}{4}\sqrt{\{\rho_{i_1 s i_2 i_3} \hat{Y}_{i_1 s}^{t,a} \hat{Y}_{i_2 i_3}^{t,a}\}_+}\Big)_{1\times \#|\tilde{\xi}_a|}}_{(i_1,i_2,i_3)\in\tilde{\xi}_a}, \underbrace{\Big(\frac{C_A}{4}\sqrt{\{-\rho_{i_1 s i_2 i_3} \hat{Y}_{i_1 s}^{t,b} \hat{Y}_{i_2 i_3}^{t,b}\}_+}\Big)_{1\times \#|\tilde{\xi}_b|}}_{(i_1,i_2,i_3)\in\tilde{\xi}_b} \Bigg),$$

$$\bar{X}_t^- = \Bigg(X_t^-, \underbrace{\Big(\frac{C_A}{4}\sqrt{\{\rho_{i_1 s i_2 i_3} \hat{Y}_{i_1 s}^{t,a} \hat{Y}_{i_2 i_3}^{t,a}\}_-}\Big)_{1\times \#|\tilde{\xi}_a|}}_{(i_1,i_2,i_3)\in\tilde{\xi}_a}, \underbrace{\Big(\frac{C_A}{4}\sqrt{\{-\rho_{i_1 s i_2 i_3} \hat{Y}_{i_1 s}^{t,b} \hat{Y}_{i_2 i_3}^{t,b}\}_-}\Big)_{1\times \#|\tilde{\xi}_b|}}_{(i_1,i_2,i_3)\in\tilde{\xi}_b} \Bigg).$$

Denote the covariance of $\bar{X}_t^+$ and $\bar{X}_t^-$ as $\bar{\Sigma}_1$ and $\bar{\Sigma}_2$. Note that each element in $\bar{X}_t^+$ or $\bar{X}_t^-$ is a bounded binary random variable. Similarly, $\bar{X}_t^+$ and $\bar{X}_t^-$ are subgaussian vectors. Therefore,

$$\boldsymbol{A_3} + \boldsymbol{B_2} = \frac{1}{M}\sum_{t=1}^{M}\big(\langle \bar{X}_t^+, \bar{X}_t^+\rangle - \langle \bar{X}_t^-, \bar{X}_t^-\rangle\big) = \frac{1}{M}(\bar{Q}_1 - \bar{Q}_2),$$

$$E(\boldsymbol{A_3} + \boldsymbol{B_2}) = \frac{1}{M}(E\bar{Q}_1 - E\bar{Q}_2).$$

Denote $s = \frac{m-l}{m}x_0 - t - \boldsymbol{A_2} - \boldsymbol{A_4} - E(\boldsymbol{A_3} + \boldsymbol{B_2})$, we estimate $E(\boldsymbol{A_3} + \boldsymbol{B_2})$, $\boldsymbol{A_2}$ and $\boldsymbol{A_4}$ in the following. Given $z_s^* = b$ and the result in (2.26), we have for some constant $c > 0$ and $q = 1, \cdots, K$:

$$E\Big[\{Y_{is}^t \log\frac{\mu_{qa}}{\mu_{qb}} + (1 - Y_{is}^t)\log\frac{1-\mu_{qa}}{1-\mu_{qb}}\}\Big] = \mu_{qb}\log\frac{\mu_{qa}}{\mu_{qb}} + (1-\mu_{qb})\log\frac{1-\mu_{qa}}{1-\mu_{qb}} < -c < 0.$$

Then

$$E\boldsymbol{A_3} = \frac{1}{M}\sum_{t=1}^{M}\sum_{q=1}^{K}\sum_{i\neq s}^{N}\Big[z_{iq}^{*}\{\mu_{qb}\log\frac{\mu_{qa}}{\mu_{qb}} + (1-\mu_{qb})\log\frac{1-\mu_{qa}}{1-\mu_{qb}}\}\Big] < -c^{*}\eta_N(N-1).$$

Given edges from different communities are independent and correlation density λ, there exists a constant $C > 0$ such that

$$\begin{aligned} E\boldsymbol{B_2} =& \frac{\rho C_A}{4}\Big[\langle\alpha_a\otimes vec(\alpha_a^T\alpha_a), E\{\hat{Y}_{\cdot s}^{t,a}\otimes vec(\hat{\boldsymbol{Y}}^{t,a})\}\rangle - \langle\alpha_b\otimes vec(\alpha_b^T\alpha_b), E\{\hat{Y}_{\cdot i}^{t,b}\otimes vec(\hat{\boldsymbol{Y}}^{t,b})\}\rangle\Big] \\ =& -\frac{\rho C_A}{4}\langle\alpha_b\otimes vec(\alpha_b^T\alpha_b), E\{\hat{Y}_{\cdot i}^{t,b}\otimes vec(\hat{\boldsymbol{Y}}^{t,b})\}\rangle \leq -C\lambda N_b^3. \end{aligned}$$

Therefore, $-E(\boldsymbol{A_3}+\boldsymbol{B_2}) \geq c'(c^{*}\eta_N N + \lambda N^3)$ for some positive constant c'. Based on condition C1 that $\mu_{ql}, q, l = 1,\cdots,K$ are bounded, it can be shown that $|EY_{is}^t\log\frac{\mu_{qa}}{\mu_{qb}} + (1-EY_{is}^t)\log\frac{1-\mu_{qa}}{1-\mu_{qb}}|$ is bounded then $|\boldsymbol{A_2}| = \mathcal{O}_N(N)$. From condition (C5), we have

$$\log\frac{\gamma_1}{\gamma_2} \leq \log\frac{\mu_{qa}\hat{\mu}_{qb}}{\hat{\mu}_{qa}\mu_{qb}} \leq \log\frac{\gamma_2}{\gamma_1} \quad \text{and} \quad \log\frac{1-\gamma_2}{1-\gamma_1} \leq \log\frac{(1-\mu_{qa})(1-\hat{\mu}_{qb})}{(1-\hat{\mu}_{qa})(1-\hat{\mu}_{qb})} \leq \log\frac{1-\gamma_1}{1-\gamma_2}$$

Define $\gamma = \max\{-\log\frac{\gamma_1}{\gamma_2}, \frac{\gamma_2}{\gamma_1}, -\frac{1-\gamma_2}{1-\gamma_1}, \frac{1-\gamma_1}{1-\gamma_2}\}$. Then we have

$$\begin{aligned} |\boldsymbol{A_4}| =& |\frac{1}{M}\sum_{t=1}^{M}\sum_{q=1}^{K}\sum_{i\neq s}^{N}\Big[\alpha_{iq}\{Y_{is}^t\log\frac{\mu_{qa}\hat{\mu}_{qb}}{\hat{\mu}_{qa}\mu_{qb}} + (1-Y_{is}^t)\log\frac{(1-\mu_{qa})(1-\hat{\mu}_{qb})}{(1-\hat{\mu}_{qa})(1-\hat{\mu}_{qb})}\}\Big]| \\ \leq& \gamma|\sum_{q=1}^{K}\sum_{i\neq s}^{N}\alpha_{iq}| \leq \gamma N \end{aligned}$$

Therefore we have $|\boldsymbol{A_2}+\boldsymbol{A_4}| = \mathcal{O}_N(N)$. We choose $t = -\frac{E(\boldsymbol{A_3}+\boldsymbol{B_2})}{2}$ and $x_0 = -\alpha' MN$ where $\alpha' > 0$. As the function of node size N, M and λ are constrained in the range $M \leq o(N^{2-\frac{\eta}{2}})$ and $\lambda N^{\frac{\eta}{2}} > 1$, where η is defined in condition C4. Then $\frac{m-l}{m}x_0 = o_N(E(\boldsymbol{A_3}+\boldsymbol{B_2}))$. Therefore, $E(\boldsymbol{A_3}+\boldsymbol{B_2})$ is dominant term in s such that $s \geq -C'\lambda N^3$ where $C' > 0$ is a constant. Follow a similar discussion in (2.29) and condition C3, we have the upper bound for $\|\bar{\Sigma}_1\|_{op}$:

$$\|\bar{\Sigma}_1\|_{op} \leq c_0(1 + c_1\lambda N^2).$$

In addition, from $\#|X_t^+| = N$, $\#|\bar{\xi}_a| = o(N_a^3)$, $\#|\bar{\xi}_b| = o(N_b^3)$ and condition (C3), we have the upper bound for $\|\bar{\Sigma}_1\|_F^2$:

$$\|\bar{\Sigma}_1\|_F^2 \leq C_1 N(1 + c_1\lambda N^2) + C_2 N^3(1 + c_2\lambda N^2),$$

where C_1, C_2, c_1, c_2 are constants. Then we estimate the upper bound for the second term in (2.36) following the similar decentralized quadratic decomposition in Theorem 2.5.1 and Theorem 2.5.3:

$$\begin{aligned}&\mathbb{P}\Big(\boldsymbol{A_3} + \boldsymbol{B_2} \geq \frac{m-l}{m}x_0 - t - \boldsymbol{A_2} - \boldsymbol{A_4}\Big) = \mathbb{P}\Big\{(\bar{Q}_1 - E\bar{Q}_1) - (\bar{Q}_2 - E\bar{Q}_2) > Ms\Big\} \\ \leq &\frac{1}{2}\mathbb{P}\Big\{|\bar{Q}_1 - E\bar{Q}_1| > \frac{Ms}{2}\Big\} + \frac{1}{2}\mathbb{P}\Big\{|\bar{Q}_2 - E\bar{Q}_2| > \frac{Ms}{2}\Big\}.\end{aligned}$$

According to the generalized Hanson-Wright inequality in ([30]):

$$\frac{1}{2}\mathbb{P}\Big\{|\bar{Q}_1 - E\bar{Q}_1| > s\Big\} \leq \exp\Big\{-C\min\big(\frac{s^2M^2}{\bar{L}_1^4\|\bar{\Sigma}_1\|_F^2\|A\|_F^2}, \frac{sM}{\bar{L}_1^2\|\bar{\Sigma}_1\|_{op}\|A\|_{op}}\big)\Big\}, \tag{2.37}$$

where $A = \boldsymbol{I}_{M\times M}$ and $\bar{L}_1$ is subgaussian norm of $\bar{X}_t^+$. Notice that

$$\frac{s^2M^2}{\bar{L}_1^4\|\bar{\Sigma}_1\|_F^2\|A\|_F^2} \geq \frac{(C'\lambda N^3)^2M^2}{\bar{L}_1^4\{C_1N(1 + c_1\lambda N^2) + C_2N^3(1 + c_2\lambda N^2)\}M}.$$

$$\frac{sM}{\bar{L}_1^2\|\bar{\Sigma}_1\|_{op}\|A\|_{op}} \geq \frac{C'\lambda N^3M}{\bar{L}_1^2c_0(1 + c_3\lambda N^2)}.$$

Given $\lambda N^{\frac{\eta}{2}} > 1$, we have for some constant $C^* > 0$

$$C\min\big(\frac{s^2M^2}{\bar{L}_1^4\|\bar{\Sigma}_1\|_F^2\|A\|_F^2}, \frac{sM}{\bar{L}_1^2\|\bar{\Sigma}_1\|_{op}\|A\|_{op}}\big) \geq C^*\lambda MN.$$

The upper bound for $\mathbb{P}\Big\{|\bar{Q}_2 - E\bar{Q}_2| > \frac{Ms}{2}\Big\}$ can be similarly obtained. Therefore,

$$\mathbb{P}\Big(\boldsymbol{A_3} + \boldsymbol{B_2} \geq \frac{m-l}{m}x_0 - t - \boldsymbol{A_2}\Big) \leq \exp(-C'\lambda MN).$$

Next, we estimate the term $\mathbb{P}\Big(\boldsymbol{A_1}+\boldsymbol{B_1}\geq t\Big)$. Notice

$$E(\boldsymbol{A_1})=E\Big[\frac{1}{M}\sum_{t=1}^{M}\sum_{q=1}^{K}\sum_{i\neq s}^{N}(\alpha_{iq}-z^*_{iq})\{Y^t_{is}-E(Y^t_{is})\}(\log\frac{\mu_{qa}}{\mu_{qb}}-\log\frac{1-\mu_{qa}}{1-\mu_{qb}})\Big]=0,$$

$$E(\boldsymbol{B_1})=\frac{\rho C_A}{4M}\sum_{t=1}^{M}\big[\langle\alpha_a\otimes vec(\alpha_a^T\alpha_a)-z^*_a\otimes vec(z^{*T}_a z^*_a),E\{\hat{Y}^{t,a}_{\cdot s}\otimes vec(\hat{\boldsymbol{Y}}^{t,a})\}\rangle-$$
$$\langle\alpha_b\otimes vec(\alpha_b^T\alpha_b)-z^*_b\otimes vec(z^{*T}_b z^*_b),E\{\hat{Y}^{t,b}_{\cdot s}\otimes vec(\hat{\boldsymbol{Y}}^{t,b})\}\rangle\big].$$

Given condition (C4) such that $\|\boldsymbol{\alpha}-z^*\|_1=cN^{1-\eta},0<\eta<1$,

$$\boldsymbol{B}_1=\frac{\rho C_A}{4M}\sum_{t=1}^{M}\big\{\langle\alpha_a\otimes vec(\alpha_a^T\alpha_a)-z^*_a\otimes vec(z^{*T}_a z^*_a),\hat{Y}^{a}_{\cdot s}\otimes vec(\hat{\boldsymbol{Y}}^{a})\rangle-$$
$$\langle\alpha_b\otimes vec(\alpha_b^T\alpha_b)-z^*_b\otimes vec(z^{*T}_b z^*_b),\hat{Y}^{b}_{\cdot s}\otimes vec(\hat{\boldsymbol{Y}}^{b})\rangle\big\}.$$

Notice that for any community $a=1,\cdots,K$,

$$\begin{aligned}\|(vec(\alpha_a^T\alpha_a)-vec(z^{*T}_a z^*_a))\|_2&\leq\|\alpha_a\otimes(\alpha_a-z^*_a)\|_2+\|(\alpha_a-z^*_a)\otimes z^*_a\|_2\\&\leq\|\alpha_a\|_2\|(\alpha_a-z^*_a)\|_2+\|(\alpha_a-Z^*_a)\|_2\|z^*_a\|_2,\\\|E(\hat{Y}^{t,a}_{\cdot s})\|_2\leq\sqrt{\frac{N}{\hat{\mu}_{aa}(1-\hat{\mu}_{aa})}},&\quad\|E(\hat{\boldsymbol{Y}}^{t,a})\|_2\leq\sqrt{\frac{N^2}{\hat{\mu}_{aa}(1-\hat{\mu}_{aa})}}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\langle\alpha_a\otimes vec(\alpha_a^T\alpha_a)-z^*_a\otimes vec(z^{*T}_a z^*_a),E\{\hat{Y}^{t,a}_{\cdot s}\otimes vec(\hat{\boldsymbol{Y}}^{t,a})\}\rangle\\\leq&\|\alpha_a\otimes vec(\alpha_a^T\alpha_a)-z^*_a\otimes vec(z^{*T}_a z^*_a)\|_2\|E\{\hat{Y}^{t,a}_{\cdot s}\otimes vec(\hat{\boldsymbol{Y}}^{t,a})\}\|_2\\\leq&\big(\|\alpha_a\otimes vec(\alpha_a^T\alpha_a)-vec(z^{*T}_a z^*_a))\|_2+\|(\alpha_a-z^*_a)\otimes vec(z^{*T}_a z^*_a)\|_2\big)\|E\{\hat{Y}^{t,a}_{\cdot s}\otimes vec(\hat{\boldsymbol{Y}}^{t,a})\}\|_2\\\leq&\|\alpha_a-z^*_a\|_2\cdot\big(\|\alpha_a\|_2^2+\|z^*_a\|_2^2+\|\alpha_a\|_2\|z^*_a\|_2\big)\cdot\|E(\hat{Y}^{t,a}_{\cdot s})\|_2\cdot\|E(\hat{\boldsymbol{Y}}^{t,a})\|_2\leq\frac{3N*N^{3/2}}{\hat{\mu}_{aa}(1-\hat{\mu}_{aa})}\|\alpha_a-z^*_a\|_2.\end{aligned}$$

Since $\|\alpha_a-z^*_a\|_2=\sqrt{\|\alpha_a-z^*_a\|_2^2}\leq\sqrt{\|\boldsymbol{\alpha}-z^*\|_1}$ for any $a=1,\cdots,K$, then for some constant

$C > 0$,

$$|E(\boldsymbol{B}_1)| \leq CN^{3-\frac{\eta}{2}}.$$

We define edge vectors $\tilde{Y}_t, t = 1, \cdots, M$ and membership vector $\boldsymbol{\theta}_{a,b}$ as:

$$\tilde{Y}_t = \big\{ \underbrace{Y^t_{\cdot s} - E(Y^t_{\cdot s}), \cdots, Y^t_{\cdot s} - E(Y^t_{\cdot s})}_{NK}, \hat{Y}^{t,a}_{\cdot s} \otimes vec(\hat{\boldsymbol{Y}}^{t,a}), \hat{Y}^{t,b}_{\cdot s} \otimes vec(\hat{\boldsymbol{Y}}^{t,b}) \big\},$$

$$\boldsymbol{\theta}_{a,b} = \Big[\underbrace{(\alpha_{iq} - z^*_{iq})(\log \frac{\mu_{qa}}{\mu_{qb}} - \log \frac{1-\mu_{qa}}{1-\mu_{qb}})}_{i=1\cdots,N}, \cdots, \underbrace{(\alpha_{iK} - z^*_{iK})(\log \frac{\mu_{Ka}}{\mu_{Kb}} - \log \frac{1-\mu_{Ka}}{1-\mu_{Kb}})}_{i=1\cdots,N},$$
$$\frac{\rho C_A}{4}\{\alpha_a \otimes vec(\alpha_a^T \alpha_a) - z^*_a \otimes vec(z^{*T}_a z^*_a)\}, \frac{\rho C_A}{4}\{\alpha_b \otimes vec(\alpha_b^T \alpha_b) - z^*_b \otimes vec(z^{*T}_b z^*_b)\} \Big].$$

Notice for $a, b = 1, \cdots, K$, we have

$$\begin{aligned}\|\boldsymbol{\theta}_{a,b}\|_2^2 &\leq \mu_2 \|\boldsymbol{\alpha} - z^*\|_2^2 + \|\alpha_a \otimes vec(\alpha_a^T \alpha_a) - z^*_a \otimes vec(z^{*T}_a z^*_a)\|_2^2 + \|\alpha_b \otimes vec(\alpha_b^T \alpha_b) - z^*_b \otimes vec(z^{*T}_b z^*_b)\|_2^2 \\ &\leq \mu_2 \|\boldsymbol{\alpha} - z^*\|_1 + C_1 N^2 (\|\alpha_a - z^*_a\|_1 + \|\alpha_b - z^*_b\|_1),\end{aligned}$$

where $\mu_2 := \max\{(\log \frac{\mu_{qa}}{\mu_{qb}} - \log \frac{1-\mu_{qa}}{1-\mu_{qb}})\},\ q = 1, \cdots, K$ and $C_1 > 0$ is a constant. Then we can transform $\mathrm{Var}(\boldsymbol{A_1} + \boldsymbol{B_1})$ into

$$\mathrm{Var}(\boldsymbol{A_1} + \boldsymbol{B_1}) = \frac{1}{M}\sum_{t=1}^{M} \mathrm{Var}(\boldsymbol{\theta}_{a,b}\tilde{Y}_t) = \frac{1}{M}\sum_{t=1}^{M} \boldsymbol{\theta}^T_{a,b} Cov(\tilde{Y}_t, \tilde{Y}_t)\boldsymbol{\theta}_{a,b} \leq \frac{1}{M}\|Cov(\tilde{Y}_t, \tilde{Y}_t)\|_{op}\|\boldsymbol{\theta}_{a,b}\|_2^2.$$

From the condition (C3) and same discussion in (2.29), we have for some constant $C > 0$ and $c > 0$:

$$\|Cov(\tilde{Y}_t, \tilde{Y}_t)\|_{op} \leq C(1 + c\lambda N^2).$$

Given $\frac{1}{\lambda} = o(N^{\frac{\eta}{2}})$, we have $E(\boldsymbol{A_1} + \boldsymbol{B_1}) = o_N(E(\boldsymbol{A_3} + \boldsymbol{B_2}))$ then the $E(\boldsymbol{A_3} + \boldsymbol{B_2})$ is dominating

in the term $\{t - E(\boldsymbol{A_1} + \boldsymbol{B_1})\}^2$. Based on the Markov inequality, for some constant $C_2 > 0$

$$\begin{aligned}
\mathbb{P}\Big(\boldsymbol{A_1} + \boldsymbol{B_1} \geq t\Big) \leq & \frac{\mathrm{Var}(\boldsymbol{A_1} + \boldsymbol{B_1})}{\{t - E(\boldsymbol{A_1} + \boldsymbol{B_1})\}^2} \leq \frac{\|Cov(\tilde{Y}_t, \tilde{Y}_t)\|_{op} \|\boldsymbol{\theta}_{a,b}\|_2^2}{M\{c'(N + \lambda N^3)\}^2} \\
\leq & \frac{C(1 + c\lambda N^2)\{\mu_2 \|\boldsymbol{\alpha} - z^*\|_1 + C_1 N^2(\|\alpha_a - z_a^*\|_1 + \|\alpha_b - z_b^*\|_1)\}}{(c'(N + \lambda N^3))^2 M} \\
\leq & \frac{2Cc\{\mu_2 \|\boldsymbol{\alpha} - z^*\|_1 + C_1 N^2(\|\alpha_a - z_a^*\|_1 + \|\alpha_b - z_b^*\|_1)\}}{c'^2(1 + \sqrt{\lambda} N^2)^2 M} \\
\leq & C_2 \frac{N^{\eta/4}(\|\alpha_a - z_a^*\|_1 + \|\alpha_b - z_b^*\|_1)}{(1 + \lambda N^{2+\frac{\eta}{4}})M}.
\end{aligned}$$

Combined upper bound of $\mathbb{P}\Big(\boldsymbol{A_1} + \boldsymbol{B_1} \geq t\Big)$ and $\mathbb{P}\Big(\boldsymbol{A_3} + \boldsymbol{B_2} \geq s\Big)$ with (2.35), there exists positive constant $c_1 > 0, c_2 > 0, c_3 > 0$ such that:

$$\begin{aligned}
& E\|\boldsymbol{\alpha}^{s+1} - \boldsymbol{z^*}\|_1 \leq 2\boldsymbol{w}NK \exp(-\alpha' MN) + 2\boldsymbol{w} \sum_{l=0}^{m-1} \sum_{a=1}^{K} \sum_{b \neq a} \sum_{i: z_i^* = b} \frac{1 - exp(\alpha' MN)}{m} E(\boldsymbol{L}_2) \\
\leq & 2\boldsymbol{w}KN \exp(-\alpha' MN) + 2\boldsymbol{w}mKN \exp(-C'\lambda MN) + 2\boldsymbol{w}mKNC_2 \frac{N^{\eta/4}(\|\alpha_a - z_a^*\|_1 + \|\alpha_b\,! - z_b^*\|_1)}{(1 + \lambda N^{2+\frac{\eta}{4}})M} \\
\leq & c_1 NK \exp(-c_2 \alpha' MN) + \frac{c_3 N^{1+\frac{\eta}{4}} \|\boldsymbol{\alpha}^s - \boldsymbol{z^*}\|_1}{(1 + \lambda N^{2+\frac{\eta}{4}})M}.
\end{aligned}$$

Chapter 3

High-order Embedding for Hyperlink Prediction

3.1 Introduction

Hyperlinks generalize the traditional pairwise links through capturing the interaction among a group of nodes. Specifically, a hyperlink is called m-size if it is a set containing m nodes. Figure 3.1 illustrates the differences between pairwise links and hyperlinks. Hyperlinks occur frequently in social networks and recommender systems involving user-driven contents. For example, websites like Delicious, Last.fm and Flickr provide online-tagging systems allowing annotations of different types of resources such as documents, music, and photos. In this case, the system involves not only pairwise relations between users and resources but also three-way user-tag-resource relations that cannot be captured by pairwise relations using the traditional network formulation. Furthermore, in sentence generation, pairwise adjacent relations between words, represented by low-order gram models, fail to capture the remote dependency and ordering among multiple words. To render cohesive contextual meaning, high-order gram models are necessary to select groups of potentially linked words.

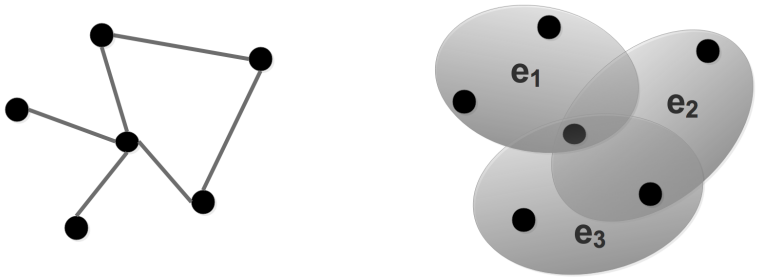
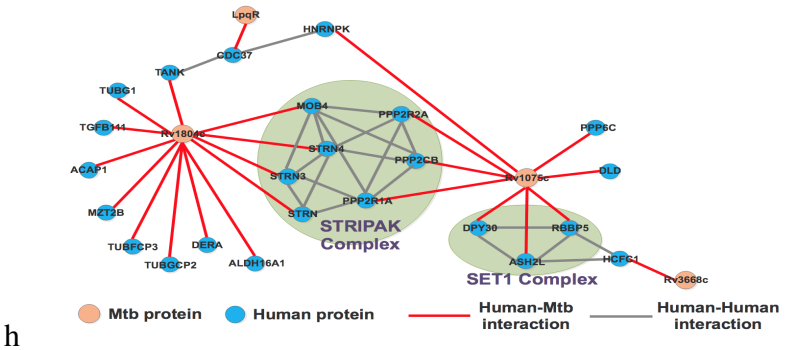


Figure 3.1: The left network is formulated by pairwise links; the right one is formulated by three hyperlinks e_1, e_2 and e_3. Each of them is a 3-size hyperlinks connecting 3 nodes.



h

Figure 3.2: Partial Mtb-human protein-protein interaction network. The human protein complexes (light green) could be formulated as hyperlinks instead of the cliques.

Another example is gene detection in a gene interaction network. The focus is to identify genes which have mutations associated with subtypes of cancer. In this situation, pairwise relations between genes fail to capture the collaborative high-order interactions of one gene with a subgroup of other genes. In other words, the traditional network formulation based on pairwise relations breaks down. However, it is essential to identify subgroups of genes which are functionally associated with each other and can potentially formulate a protein complex [36, 99, 35, 42]. .From a biological perspective, many other networks typically involve multilateral and high-order relations [61] in order to reflect the complex relations among proteins and metabolism. Figure 3.3 illustrates part of interaction network between Mycobacterium tuberculosis (Mtb) proteins and human proteins. In general, Mtb proteins interact with the host proteins and replicate inside of host's immune cells. Through introducing the hyperlink for the human protein complex instead of modeling human protein individually, we are able to capture the interaction of Mtb protein with human protein complex, while the traditional pairwise link representation does not have the capacity to achieve this goal.

As indicated by these examples, it is critical to identify the potential multiway relations among multiple units, or high-order relations represented by hyperlinks. Unfortunately, existing methods for prediction and inference of hyperlinks use the information only from observed pairwise links [52, 55, 107, 6, 76, 72, 73, 5] or observed hyperlinks [61, 71, 125, 2, 123, 124, 46]. One major challenge for these methods is that hyperlinks are likely to be partially observed or completely missing in practice. Therefore, unlike the case of pairwise relations, inferring high-order relations is much more challenging and requires additional modeling effort. For example, identification

of a hyperlink of a protein complex connecting with multiple genes requires additional biological experimentation and validation, while, in social networks, inferences about local social circles may require additional information involving high-order interactive relations.

Related works on classification and community detection require an inference of hyperlinks from pairwise links. For instance, [32, 95, 2] suggest methods based on hyperlink expansions or random walks to reconstruct hyperlinks from pairwise links. These methods use a principle of generating hyperlinks based on pre-specified relations among hyperlinks and pairwise links, while treating hyperlinks as a subgraph with a certain configuration such as a fully-connected clique or star-shaped subgroup of nodes. However, this heuristic principle is not feasible to uncover complex network structures defined by high-order relations. Moreover, they are not adaptive to different structures of hyperlinks, which tend to lead misspecified hyperlinks, especially in the absence of the true knowledge of pairwise links and hyperlinks. In addition, the aforementioned methods mainly focus on node classifications and detection of the global community structures rather than identifying subgroup structures.

To overcome the foregoing difficulties of modeling hidden structures of high-order relations, in this chapter we develop a novel network embedding procedure to jointly model pairwise links and high-order hyperlinks simultaneously to capture complex high-order interactions. In particular, we proceed in a hierarchical fashion in that pairwise and multiway relations are modeled at different resolutions. E.g., pairwise relations are structured by low-level node-wise network features, while hyperlinks capture multiway relations via high-level subgroup-wise features. Jointly they can identify the subgroup configuration and capture the hyperlink-generating features effectively. This empowers incorporating more complete and rich information from the observed network. One advantage of this hierarchical modeling is that hyperlink prediction can borrow information from the existing known pairwise relations. More specifically, in the presence of a hyperlink connecting two nodes, they are more likely to form a pairwise relation as compared with nodes in the absence of a hyperlink. On the other hand, nodes that are highly connected by pairwise links may suggest the presence of a potential hyperlink among them. In addition, this principle of network connectivity also reflects the nature of reality. For example, any two proteins connected by a hyperlink defined as a protein complex [41, 98, 61], have a higher probability to build a pairwise connection; while a group of proteins with a dense or certain pattern of pairwise functional associations is

more likely to form a hyperlink as a functional subgroup [13, 94, 59, 80]. In general, the proposed hyperlink prediction framework provides cohesive statistical modeling for both pairwise links and hyperlinks, which enhance the mutual inference between pairwise links and hyperlinks.

In addition, we provide the consistency of the proposed estimation and show that our method achieves a faster convergence rate compared to the existing embedding procedures only utilizing pairwise link information since we are able to incorporate either the observed or inferred hyperlinks via the joint embedding procedure. The theoretical development using the large deviation theory is nontrivial, since the pairwise links and hyperlinks are correlated intrinsically in that the independent model cannot be assumed. In contrast, the most of the existing probability concentration tools are established under the independent model. Furthermore, the established theoretical properties can be extended to joint embedding procedure with a general loss function beyond the L_2 loss in this paper, hence more complex and non-linear latent features can be captured.

This chapter is organized as follows: Section 3.2 introduces the background and necessary notations of the proposed method. Section 3.3 introduces the proposed joint embedding method and the inference procedure from observed pairwise link to hyperlinks. Section 3.4 illustrates the theoretical properties of the proposed embedding method under the scenarios where the hyperlinks are observed or inferred from pairwise links. Section 3.5 demonstrates simulation studies, and Section 3.6 illustrates an application of proposed method on the Facebook ego-network. The last section provides conclusions and some further discussion.

3.2 Background and Notations

We define an observed network $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} = \{v_i\}_{i=1}^N$ denotes a set of N nodes and $\boldsymbol{E}$ is a set of observed pairwise links. For an undirected and unweighted network, $\boldsymbol{G}$ can be represented through a symmetric binary adjacent matrix $\boldsymbol{Y} = \{Y_{ij}\}_{1\le i\neq j\le N}$ in that $Y_{ij} = 1$ if $e_{ij} \in E$, otherwise $Y_{ij} = 0$. In addition, we define an m-order uniform hypergraph on $\boldsymbol{V}$ as $\boldsymbol{G_H} = (\boldsymbol{V}, \boldsymbol{H} = \{e_{\boldsymbol{i}}\}_{\boldsymbol{i}\in I})$, where I is an index set of m-tuple indices $\boldsymbol{i} = \{i_1, i_2, \cdots, i_m\}$. To represent the hyperlink, we introduce a m-order tensor $\mathcal{Y} \in \{0,1\}^{R^m}$ such that $Y_{i_1 i_2,\cdots,i_m} = 1$ if there exists an m-order hyperlink connecting nodes $v_{i_1}, v_{i_2}, \cdots, v_{i_m}$, i.e., $e_{i_1,i_2,\cdots,i_m} \in \boldsymbol{H}$ and $Y_{i_1 i_2,\cdots,i_m} = 0$ otherwise. Denote the sets of observed pairwise links and hyperlinks as $\Omega_{\boldsymbol{Y}}$ and $\Omega_{\mathcal{Y}}$.

Therefore, the number of observed pairwise links and hyperlinks are $|\Omega_{\boldsymbol{Y}}|$ and $|\Omega_{\mathcal{Y}}|$.

We develop a generative learning framework for hyperlink prediction. Specifically, we assume that both the pairwise link Y_{ij} and hyperlinks $Y_{i_1 i_2,\cdots,i_m}$ are generated through Bernoulli distribution through the interaction among their endpoint vertices as $\boldsymbol{P}\big(Y_{ij}|v_i, v_j\big)$ and $\boldsymbol{P}\big(Y_{i_1 i_2,\cdots,i_m}|v_{i_1}, \cdot\cdot, v_{i_m}\big)$.

The proposed method for hyperlink prediction consists of two steps. First, for each node v_i, a r-dimensional latent vector representation $Z_i = (Z_{i1}, \cdots, Z_{ir})$ is introduced so that the concordance between Z_i and $\{Z_j\}_{j\neq i}$ represents the observed or inferred relations between node v_i and other nodes $\{v_j\}_{j\neq i}$, where each element in Z_i represents a latent feature of node v_i and encodes its local neighborhood information of a network. Typically, r needs to be much smaller than N because mapping an observed network onto a latent space would increase the estimation efficiency of latent factors while reducing the variation of prediction. In the second step, underlying pairwise links or hyperlinks are inferred based on estimated latent feature factors $\{Z_i\}_{i=1}^N$ of each node. In general, there is a high probability that nodes are connected through a hyperlink when their corresponding latent factors have a strong concordance.

3.3 Methodology

3.3.1 Reconstruction of Pairwise Relations in a Latent Space:

To better understand the key idea, we begin our discussion with pairwise relations. To incorporate observed pairwise relations into a latent space, we propose to minimize a cost function to estimate latent factors $\boldsymbol{Z} = \{Z_i\}_{i=1}^N$:

$$L_1(\boldsymbol{Z}) = -\frac{1}{|\Omega_{\boldsymbol{Y}}|}\sum_{Y_{ij}\in\Omega_{\boldsymbol{Y}}}(Y_{ij} - \log\sigma\Big[Z_i Z_j^T\Big])^2, \tag{3.1}$$

where $|\Omega_{\boldsymbol{Y}}|$ is the number of total observed pairwise links, β_1 is the offset parameters and $\sigma(\cdot)$ is a link function that transforms the concordance between latent factors from two nodes (measured by a larger value of $Z_i Z_j^T$) into the probability of a pairwise link, e.g, a sigmoid function as we adopted in this paper. Intuitively, two latent factors Z_i and Z_j are encouraged to be similar in the presence of a pairwise link between nodes v_i and v_j with $Y_{ij} = 1$. Otherwise, latent factors from

two isolated nodes tend to be dissimilar.

3.3.2 Reconstructing Inferred High-order Relations in Latent Space:

Hyperlink prediction is more challenging since hyperlinks are often unobserved, and configurations for high-order associations involve more uncertainty and requires more sophisticated high-dimensional modeling tools. One key innovation of the proposed method is to infer hyperlinks using the structure of latent factors expressed in high-order tensors. Assume that m-order relations are defined as m-order hyperlinks, we can formulate them by an m-order tensor $\mathcal{Y} \in \{0,1\}^{N^m}$. Specifically, if there is a hyperlink connecting m nodes $\{v_{i_1}, \cdots, v_{i_m}\}$, then $Y_{i_1 i_2 \cdots i_m} = 1$. To reduce the dimensionality, we model the m-order hyperlinks as a low-rank concordance structure on the latent feature space of nodes $\boldsymbol{Z}$ through the CANDECOMP/PARAFAC (CP) tensor decomposition:

$$P(Y_{i_1 i_2, \cdots, i_m} = 1) = \sigma\Big(\sum_{(i,j) \in \{i_1 i_2, \cdots, i_m\}} Z_i Z_j^T + \Big[\sum_{k=1}^{r} Z_{i_1 k} Z_{i_2 k} \cdots Z_{i_m k} \Big] \Big)$$

. Analogous to measuring the pairwise concordance via an inner product among latent factors, we apply the generalized inner product $\sum_{k=1}^{r} Z_{i_1 k} Z_{i_2 k} \cdots Z_{i_m k}$ to measure the joint concordance among a group of m latent factors. Note that the joint concordance cannot be directly inferred by the pairwise concordance in the sense that even $Z_i Z_j^T$ is large for any pair $(i,j) \in \{i_1, i_2, \cdots, i_m\}$, though it is still possible that $\sum_{k=1}^{r} Z_{i_1 k} Z_{i_2 k} \cdots Z_{i_m k}$ is small. This implies that the joint concordance capturing high-order relations cannot be substituted by pairwise relations. In addition, incorporating both pairwise links and hyperlinks on the same latent space also introduces a dependency between hyperlinks and pairwise links as $\sum_{k=1}^{r} Z_{i_1 k} Z_{i_2 k} \cdots Z_{i_m k}$ is correlated with $\{Z_i Z_j^T\}_{(i,j) \in \{i_1, \cdots, i_m\}}$, implying that we should utilize the dependency for hyperlink and pairwise link prediction as they can borrow information from each other. More importantly, a hierarchical dependency allows a better interpretation in many scientific problems as compared to performing inferences of pairwise links or hyperlinks separately.

Subsequently, we incorporate the inferred hyperlinks information into the latent space of nodes,

and encourage the latent factors $\boldsymbol{Z}$ such that the below hyperlink loss function decreases:

$$L_2(\boldsymbol{Z}) = -\frac{1}{|\Omega_{\mathcal{Y}}|} \sum_{Y_{i_1 i_2 \cdots i_m} \in \Omega_{\mathcal{Y}}} w_{i_1 i_2 \cdots i_m} \Big\{ Y_{i_1 i_2 \cdots i_m} - \sigma\big(\sum_{(i,j) \in \{i_1 i_2, \cdots, i_m\}} Z_i Z_j^T + \sum_{k=1}^{r} Z_{i_1 k} Z_{i_2 k} \cdots Z_{i_m k} \big) \Big\}, \tag{3.2}$$

where $w_{i_1 i_2, \cdots, i_m}$ is the weight for observed or inferred hyperlink $Y_{i_1 i_2 \cdots i_m}$, $\Omega_{\mathcal{Y}}$ is the set of incorporated m-order hyperlinks, and $|\Omega_{\mathcal{Y}}|$ is the total number of hyperlinks. If there exists a potential hyperlink connecting nodes $\{i_1, i_2, \cdots, i_m\}$ in a hypergraph, then decreasing the loss function in (3.2) encourages a joint concordance among latent factors $\{Z_{i_1}, Z_{i_2}, \cdots, Z_{i_m}\}$. Consequently, we preserve inferred high-order relations among nodes in the embedding latent space of nodes. In terms of latent factors estimation, (3.1) and (3.2) serve as the second-order moment and the m-order moment estimations of $\boldsymbol{Z}$, respectively. Intuitively, incorporating additional moment information of latent factors $\boldsymbol{Z}$ reduces the estimation bias while increasing the efficiency, which leads to a more accurate estimation of latent factor $\boldsymbol{Z}$.

3.3.3 Joint Network Embedding for Pairwise Links and Hyperlinks Prediction:

Combining the previous two parts, we estimate the latent features of nodes by jointly incorporating observed pairwise links and inferred hyperlinks through the following combined loss function:

$$\begin{aligned} L(\boldsymbol{Z}) = & -\frac{1}{|\Omega_{\boldsymbol{Y}}|} \sum_{Y_{ij} \in \Omega_{\boldsymbol{Y}}} (Y_{ij} - \sigma\Big[Z_i Z_j^T\Big])^2 \\ & - \frac{1}{|\Omega_{\mathcal{Y}}|} \sum_{Y_{i_1 i_2 \cdots i_m} \in \Omega_{\mathcal{Y}}} w_{i_1 i_2 \cdots i_m} \Big\{ Y_{i_1 i_2 \cdots i_m} - \sigma\big(\sum_{(i,j) \in \{i_1 i_2, \cdots, i_m\}} Z_i Z_j^T + \sum_{k=1}^{r} Z_{i_1 k} Z_{i_2 k} \cdots Z_{i_m k} \big) \Big\}^2 \\ & + \lambda \|\boldsymbol{Z}\|^2, \end{aligned} \tag{3.3}$$

Due to randomness or noisy sources of information, spurious pairwise links can cause some hyperlinks to be incorrectly observed or inferred. Therefore, we impose the weight function $w_{i_1 i_2, \cdots, i_m}$ in (3.2) to downweigh these spurious links through penalization. In addition, in a network system with many potential hyperlinks, we are more interested in those important hyperlinks in the sense

that they either capture the local subgroup structures in the network or have high importance. To achieve these two goals, we adopt the penalization term $\|\boldsymbol{Z}\|^2$ to control the total concordance among the latent features of nodes such that we penalize more on those isolated hyperlinks which are highly discordant with observed pairwise links. Furthermore, the penalization term $\|\boldsymbol{Z}\|^2$ imposes a low-rank structure of latent features which can mitigate the singularity problem when the degree of nodes is smaller than the rank of latent factors.

3.3.4 Inferring Potential Hyperlinks through Observed Pairwise Links

Another innovation of the proposed framework is that potential unobserved hyperlinks can be inferred from observed pairwise links. Although desired high-order relations represented by hyperlinks can be captured by the complex subgroup structure of a network, only a small number of hyperlinks are directly observed to infer potential hyperlinks in many applications. Nevertheless, it is still feasible to infer potential hyperlinks from observed pairwise links through their dependent information. This is due to the fact that pairwise links and hyperlinks characterize similar types of relations but at different group levels. In the proposed framework, this dependency is introduced through node-wise latent factors in a hierarchical fashion, serving as building blocks for modeling concordance shared by both pairwise links and hyperlinks. Therefore, we can incorporate the inferred hyperlink information to capture the subgroup structures. For illustration, we state the inference procedure for a three-order hyperlink, which consists of two steps based on the hyperlink modeling in (3.2):

$$P(Y_{i_1 i_2 j_3} = 1) = \sigma\big(Z_{i_1}^T Z_{i_2} + Z_{i_1}^T Z_{i_3} + Z_{i_2}^T Z_{i_3} + \sum_{k=1}^{r} Z_{i_1 k} Z_{i_2 k} Z_{i_3 k}\big). \tag{3.4}$$

First, we approximate the pairwise concordance $Z_{i_1}^T Z_{i_2} + Z_{i_1}^T Z_{i_3} + Z_{i_2}^T Z_{i_3}$ within the potential hyperlink $Y_{i_1 i_2 j_3}$. In the second step, the three-order concordance is approximated through the similarity of neighbourhood among nodes $\{i_1, i_2, i_3\}$. Next, we assign hyperlinks to those three-way tuples $\{i_1, i_2, i_3\}$ that have large pairwise concordance and three-order concordance simultaneously.

Step 1: Approximate pairwise concordance within hyperlinks: Given that the probability of pairwise links Y_{ij} depends on the pairwise concordance $Z_i^T Z_j$ based on (3.1), it is reasonable

to approximate $Z_{i_1}^T Z_{i_2} + Z_{i_1}^T Z_{i_3} + Z_{i_2}^T Z_{i_3}$ through $Y_{i_1 i_2} + Y_{i_1 i_3} + Y_{i_2 i_3}$. Define the set of node combination $\Omega_1 = \{(i_1, i_2, i_3) | Z_{i_1}^T Z_{i_2} + Z_{i_1}^T Z_{i_3} + Z_{i_2}^T Z_{i_3} \text{ is large}\}$. Therefore, we can approximate Ω_1 by a set $\hat{\Omega}_1 = \{(i_1, i_2, i_3) | Y_{i_1 i_2} + Y_{i_1 i_3} + Y_{i_2 i_3} \geq \eta_1\}$ with a specified positive threshold η_1, where η_1 can be chosen as the empirical quantile of $\{Z_i^T Z_i\}_{1 \leq i < j \leq N}$. Intuitively, if all the nodes within a hyperlink are pairwisely connected, then their latent features are close in terms of the distance on the latent space spanned by $\boldsymbol{Z}$, thus lead to a large value for $Z_{i_1}^T Z_{i_2} + Z_{i_1}^T Z_{i_3} + Z_{i_2}^T Z_{i_3}$.

Step 2: Approximate three-order concordance within hyperlinks: Define the set of node combination $\Omega_2 = \{(i_1, i_2, i_3) | \sum_{k=1}^r Z_{i_1 k} Z_{i_2 k} Z_{i_3 k} \text{ is large}\}$. The main difference between the pairwise concordance and three-order concordance is that former indicates latent features shared by pairwise nodes while the latter indicates the latent features shared by all three nodes. To account for the high-order similarity, we consider the similarity between node i and j as $corr(Y_{\cdot i}, Y_{\cdot j})$, measuring the global similarity between node i and j in terms of their neighbourhood pattern. A large $corr(Y_{\cdot i}, Y_{\cdot j})$ indicates Z_i and Z_j are more concordant in each element. Therefore, we approximate Ω_2 by the set:

$$\hat{\Omega}_2 = \{(i_1, i_2, i_3) | corr(Y_{\cdot i_1}, Y_{\cdot i_2}) + corr(Y_{\cdot i_1}, Y_{\cdot i_3}) + corr(Y_{\cdot i_2}, Y_{\cdot i_3}) \geq \eta_2\}$$

with specific positive threshold η_2, e.g., an empirical quantile, where Ω_2 collects those nodes such that their latent factors are more similar.

Finally, we collect a 3-tuples (i_1, i_2, i_3) of nodes such that $(i_1, i_2, i_3) \in \hat{\Omega}_1 \cap \hat{\Omega}_2$ among which have a higher probability to formulate a three-way hyperlink according to (3.4). Therefore, $\hat{\Omega}_1 \cap \hat{\Omega}_2$ is treated as inferred hyperlinks and can be incorporated into the joint embedding loss function in (3.3).

3.3.5 Embeddings Estimation

We embed each node into the latent features $\boldsymbol{Z}$ which is estimated by minimizing the joint loss function in (3.3). In contrast to existing pairwise link or hyperlink embedding approaches involving matrix factorization [76, 96, 3, 19, 28, 88, 123, 124], the proposed embedding approach utilizes tensor decomposition to preserve high-order relations in the latent space, which could entail high computational cost especially when the order of tensor m is high. However, since there is no order-

ing for the hyperlinks connecting nodes, the inferred hyperlinks tensor $\mathbf{Y}$ is super-symmetric such that $Y_{i_1 i_2,\cdots,i_m} = Y_{\varphi(i_1)\varphi(i_2),\cdots,\varphi(i_m)}$, where φ is the order permutation mapping. On this ground, we develop the following algorithm to minimize (3.3) while taking advantage of the super-symmetry of a hyperlink tensor to reduce the computational cost.

In general, we estimate the embedding of nodes $\boldsymbol{Z}$ through the coordinate gradient descent algorithm where both the gradient and Hessian matrix have explicit forms. The detailed algorithm is summarized as follows:

Algorithm: Gradient Descent Algorithm with Parallel Computing

---

1. *(Initialization)* Input observed pairwise links $\boldsymbol{Y}$, inferred hyperlinks $\mathcal{Y}$, hyperlinks weights $\mathbf{W}$, the rank r, tuning parameters λ, the learning rate η, the initial value $\boldsymbol{Z}^{(0)}$ and the error bound $\varepsilon > 0$.

2. *(Latent-vectors $\boldsymbol{Z}$ update)* At the sth iteration $(s \geq 1)$, update $\boldsymbol{Z}^{(s)}$.

   (i) Update each $Z_i^{(s)}$; $i = 1, \cdots, N$, iteratively using a gradient descent formula:

$$Z_i^{(s+1)} = Z_i^{(s)} - \eta \Big[\frac{\partial^2 L(Z_i; \boldsymbol{Z}^{(s)})}{\partial Z_i^2}\Big]^{-1} \frac{\partial L(Z_i; \boldsymbol{Z}^{(s)})}{\partial Z_i}, \tag{3.5}$$

   where $\frac{\partial L(Z_i;\boldsymbol{Z}^{(s)})}{\partial Z_i}$, $\frac{\partial^2 L(Z_i;\boldsymbol{Z}^{(s)})}{\partial Z_i^2}$ are the first and second derivatives in terms of Z_i.

3. *(Stopping Criterion)* Terminate if $\frac{|L(\boldsymbol{Z}^{(s)}) - L(\boldsymbol{Z}^{(s-1)})|}{L(\boldsymbol{Z}^{(s-1)})} < \varepsilon$. Set $\hat{\boldsymbol{Z}} = \boldsymbol{Z}^{(s)}$. Otherwise set $s \leftarrow s + 1$ and go to step 2.

One advantage of the proposed algorithm is that it utilizes the super-symmetry property of a hyperlink tensor to update the latent vector corresponding to different tensor modes simultaneously instead of updating each mode iteratively. In addition, the node-wise latent vector updating in (3.5) can be performed independently of each other, which makes it feasible for parallelization to accelerate computation.

3.3.6 Link Prediction through Node-wise Embedding:

After mapping each node into the latent space spanned by column vectors of $(Z_1^T, Z_2^T, \cdots, Z_N^T)$ consisting of the observed pairwise links and inferred hyperlinks, we predict potential pairwise links and hyperlinks through an estimated degree of concordance among the latent features of nodes. Specifically, to predict a pairwise link between nodes v_i and v_j, consider

$$P\big(Y_{ij} = 1|(v_i, v_j)\big) = \exp(Z_i Z_j^T)/(1 + \exp(Z_i Z_j^T)). \tag{3.6}$$

Similarly, to predict an m-order hyperlink among a group of nodes $v_{i_1}, v_{i_2}, \cdots, v_{i_m}$, we have

$$P\big(Y_{i_1 i_2, \cdots, i_m} = 1|(v_{i_1}, v_{i_2}, \cdots, v_{i_m})\big) = \sigma\Big[\sum_{(i,j)\in\{i_1 i_2, \cdots, i_m\}} Z_i Z_j^T + \sum_{k=1}^{r} Z_{i_1 k} Z_{i_2 k} \cdots Z_{i_m k}\Big]. \tag{3.7}$$

Although the main focus is link prediction, identifying the latent feature space of nodes is also fundamentally important as it permits exploration of other types of network structures. For example, in community detection, detection of the community structures of a network bear consequences in biology, marketing, and social science. In other situations, we may develop a clustering algorithm to identify homogeneous subgroups of nodes in a latent space for an embedding set of learned nodes. Moreover, node classification can be performed in a semi-supervised fashion to predict a node's label, in which only a small proportion of nodes are labeled while a large proportion of nodes are unlabeled.

One direct application of the proposed method is document categorization. First, each word in a document is projected as a latent vector representation according to word co-occurrences in a dictionary, where a pairwise link means their adjacent relation between two words in a document and a hyperlink indicates joint semantic similarity among a group of words. Then the partial observed labels of words and their embedding representations are used for a downstream classifier for categorization.

3.4 Theoretical Results

In this section, we focus on establishing the theoretical guarantee for the proposed joint embedding methods. Specifically, we obtain the asymptotic properties of the predicted link generating probability as it directly associate with the prediction accuracy. Consider the underlying pairwise link generating process associated with the node-wise latent factors $\boldsymbol{Z}$:

$$E(Y_{ij}) = \theta_{ij}^{pair} = \frac{\exp(Z_i Z_j^T)}{1+\exp(Z_i Z_j^T)};\ 1 \le i < j \le N. \tag{3.8}$$

The hyperlink generating process associates with the latent factors $\boldsymbol{Z}$ through:

$$E(Y_{i_1 i_2 \cdots i_m}) = \theta_{i_1 i_2 \cdots i_m}^{hyper} = \sigma\Big(\sum_{1\le i<j\le m} Z_i Z_j^T + \sum_{k=1}^{r} Z_{1r} Z_{2r} \cdots Z_{mr}\Big); \tag{3.9}$$

$$\text{where}\ \ \sigma(\cdot) = \exp(\cdot)/(1+\exp(\cdot)),\ 1 \le i_1 < i_2 < \cdots < i_m \le N.$$

Denote the parameters set $\boldsymbol{\Theta} = \{\theta_{ij}^{pair}, \theta_{i_1 i_2 \cdots i_m}^{hyper};\ 1 \le i \ne j \le N,\ 1 \le i_1 \ne \cdots \ne i_m \le N\} \in \mathcal{S} \subseteq R^{N\times N} \cup R^{N^m}$. Given the generating probability set Θ, either hyperlinks or pairwise links are generated independently from the Bernoulli distribution $Bern\Big\{1, P\big\{\sigma(\boldsymbol{\Theta})\big\}\Big\}$, where $\sigma(\cdot)$ is the logistic link function. Therefore the prediction accuracy only relies on the estimation error of $\boldsymbol{\Theta}$, and the link prediction accuracy can be established through investigating the convergence property of $\boldsymbol{\Theta}$ estimator.

In the following, we establish the asymptotic property of the proposed estimator solving the joint embedding loss function:

$$l_{joint}(\boldsymbol{\Theta}; \boldsymbol{Y}, \mathcal{Y}) = l_{pair}(\boldsymbol{\Theta}; \boldsymbol{Y}) + l_{hyper}(\boldsymbol{\Theta}; \mathcal{Y}) + \lambda \|\boldsymbol{\Theta}\|^2, \tag{3.10}$$

where $l_{pair}(\boldsymbol{\Theta}; \boldsymbol{Y})$ and $l_{hyper}(\boldsymbol{\Theta}; \mathcal{Y})$ are the loss functions in (3.1) and (3.2) representing the embedding either through pairwise links or hyperlinks. Consider a sample estimator $\hat{\boldsymbol{\Theta}}$ satisfying:

$$l_{joint}(\hat{\boldsymbol{\Theta}}; \boldsymbol{Y}, \mathcal{Y}) \le \inf_{\Theta \in \mathcal{S}} l_{joint}(\boldsymbol{\Theta}; \boldsymbol{Y}, \mathcal{Y}) + \tau, \tag{3.11}$$

where τ goes to zero as the number of observed links increases. Because of the non-convex nature

of loss function (3.10) in terms of $\mathbf{\Theta}$, obtaining the global minimizer for (3.10) is in general challenging. However, we establish the convergence property for an alternative estimator satisfying (3.11) such that it only needs to be approaching the global minimizer as the sample size increases. In practice, we can seek a suboptimal solution with less computational cost instead of the optimal solution which can be either infeasible or computationally expensive. To establish the estimation consistency, we have the following assumption:

(C1): the node-wise latent factors are uniformly bounded such that $\|\boldsymbol{Z}\|_\infty \leq C$ for some positive constant C. Therefore the parameter set $\mathbf{\Theta}$ is also uniformly bounded.

Remark 3.1.: This assumption requires that the underlying search space for the latent factors is bounded which is a common assumptions for the latent factor model.

In the following, we establish the link prediction consistency through the convergence property of the sample estimator $\hat{\mathbf{\Theta}}$ based on observed pairwise links and observed hyperlinks.

Theorem 3.1. *Denote that $\mathbf{\Theta}_0$ is the underlying true link generating probability. Given the assumption (C1), for a sample estimator $\hat{\mathbf{\Theta}}$ satisfying (3.11), we have:*

$$P\left(\frac{\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0\|_F}{\sqrt{N^3+N^2}} \geq \eta\right) \leq 7\exp\left(-c(|\Omega_{\boldsymbol{Y}}| + |\Omega_{\mathcal{Y}}|)\eta^2\right),$$

where $c \geq 0$ is a constant, $\eta = \max\left(\varepsilon, \lambda^{1/2}\right)$, and the best possible rate $\varepsilon \sim \left(\frac{1}{(|\Omega_{\boldsymbol{Y}}|+|\Omega_{\mathcal{Y}}|)^{1/2}}\right)$ is achieved when $\lambda \sim \varepsilon^2$ with $\sim$ denoting shrinking at the same order.

Theorem 3.1 states that if the magnitude of penalty term shrinks to zero with an appropriate rate as the sample size of links $|\Omega_{\boldsymbol{Y}}| + |\Omega_{\mathcal{Y}}|$ increases, then the proposed method can achieve the convergence rate of $\frac{1}{(|\Omega_{\boldsymbol{Y}}|+|\Omega_{\mathcal{Y}}|)^{1/2}}$ at most. The Theorem 3.1 provides a theoretical guarantee for the proposed joint embedding strategy in the sense that the estimator converges faster compared to the estimator utilizing only pairwise links $\boldsymbol{Y}$ or hyperlinks $\mathcal{Y}$, which correspond to rate $\frac{1}{|\Omega_{\boldsymbol{Y}}|^{1/2}}$ and $\frac{1}{|\Omega_{\mathcal{Y}}|^{1/2}}$ respectively. Intuitively, both observed pairwise links and hyperlinks serve as independent sample moments with different orders. Therefore, the proposed method is able to utilize all the sample information to achieve a faster convergence.

Next we develop the convergence property of the proposed joint embedding estimator that incorporates the inferred hyperlinks instead of directly observed hyperlinks. Intuitively, the hierarchical

dependency between pairwise links and hyperlinks allows us to infer unobserved potential hyperlinks from the observed pairwise links, and thus to recover partial high-order relations from the second-order relation. The main difference between incorporating observed hyperlinks and inferred hyperlinks is that the benefits of inferred hyperlinks depend on the quality of the inferred procedure. In this paper, we consider the case for inferring the 3-order hyperlink and establish the result accordingly. For a general m-order hyperlink case, the theoretical justification is analog to the third-order hyperlink case, but requires more intensive analysis as each hyperlink is correlated with more pairwise links.

Consider an inference procedure $\mathcal{M}(\cdot) \in [0, 1]$:

$$Y_{i_1 i_2 i_3} = \mathcal{M}(Y_{i_1 i_2}, Y_{i_1 i_3}, Y_{i_2 i_3}, \boldsymbol{Y})$$

where $(Y_{i_1 i_2}, Y_{i_1 i_3}, Y_{i_2 i_3})$ indicates pairwise connection status among nodes $\{i_1, i_2, i_3\}$, and $Y_{i_1 i_2 i_3}$ is corresponding the inferred hyperlink connection status. To establish the convergence property, it is necessary to bound the inference error through the following assumption:
(C2): For a third-order hyperlink inference procedure $\mathcal{M}(\cdot)$, and $\varepsilon > 0$, we assume that

$$\frac{1}{|\Omega_{\mathcal{Y}}|} \sum_{Y_{i_1 i_2 i_3} \in \Omega_{\mathcal{Y}}} \left(\mathcal{M}(Y_{i_1 i_2}, Y_{i_1 i_3}, Y_{i_2 i_3}) - Y_{i_1 i_2 i_3}\right)^2 \leq O(\varepsilon^2),$$

where $Y_{i_1 i_2 i_3}$ is a hyperlink given the underlying true generating probility $\theta_{i_1 i_2 i_3}$ and ε is the bias for the inferred hyperlinks. Since the effective sample size of inferred hyperlinks depends on the number of observed pairwise links, assumption (C2) indicates that the average inference bias should decrease as the number of observed pairwise links increases.

In Theorem 3.2, we establish the convergence property for the joint embedding estimator via incorporating inferred hyperlinks:

Theorem 3.2. *Assume that the third-order hyperlinks are inferred through an inference procedure satisfying assumptions (C1) and (C2) with $\varepsilon > 0$, then for an sample estimator $\hat{\boldsymbol{\Theta}}$ satisfying (3.11), we have:*

$$P\left(\frac{1}{\sqrt{N^3 + N^2}} \|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \geq \eta\right) \leq 7 \exp\left(-c_1 |\Omega_{\boldsymbol{Y} \cup \hat{\boldsymbol{Y}}}| \eta^4\right),$$

where c_1 *is positive constant, and* $|\Omega_{\boldsymbol{Y}\cup\hat{\boldsymbol{Y}}}| = \frac{|\Omega^2_{\boldsymbol{Y}}|/N+|\Omega_{\boldsymbol{Y}}|}{\max_i d_i}$*, and* d_i *denotes the degree of node* $i \in \{1, 2, \cdots, N\}$*,* $\eta = \max\left(\varepsilon, \lambda^{1/2}\right)$*. The best possible rate* $\varepsilon \sim \frac{1}{|\Omega_{\boldsymbol{Y}\cup\hat{\boldsymbol{Y}}}|^{1/2}}$ *can be achieved when* $\lambda \sim \varepsilon^2$.

Remark 3.2.: The consistency established in Theorem (3.1) and Theorem (3.2) are readily to be generalized for a broad class of loss function. Specifically, the convergence rate is determined by the metric entropy ([87]) of the parameters space, which depends on the quantity $\omega = \frac{\alpha}{rN}$ with α denoting the smoothness of loss function, and rN being the number of total parameters. In this paper, since we adopt the L_2 loss function and the logistic link function to connect underlying parameters Θ and link generating probabilities in (3.8) and (3.9), then $\omega = \infty$ as both L_2 loss function and logistic function are infinite differentiable. For the general loss function, the convergence rate can be determined through estimating the corresponding metric entropy following Theorem 5.2 of ([23]).

Instead of utilizing any observed high-order information directly, the proposed method augments the sample size $|\Omega_{\boldsymbol{Y}}|$ from observed pairwise links to the size $\frac{|\Omega^2_{\boldsymbol{Y}}|/N+|\Omega_{\boldsymbol{Y}}|}{\max_i d_i}$ through the inferring hyperlinks. The data augmentation relies on the hierarchical dependency between pairwise links and hyperlinks, and Theorem 3.2 states that the proposed joint embedding integrating the data augmentation leads to a faster convergence rate the ones embedding only through observed sample $\Omega_{\boldsymbol{Y}}$ if the bias from inferring procedure can be controlled. However, the extent of the data augmentation from pairwise links could be limited by the set of sample pairwise links since the inferred hyperlinks could be non-informative on the unobserved pairwise links. In addition, the inference naturally introduces the dependency between pairwise links and hyperlinks, or dependency between hyperlinks and hyperlinks sharing overlapped nodes, which reduces the effect size of augmented hyperlinks. This is rather different to the observed hyperlink case where all pairwise links and hyperlinks are independent samples.

3.5 Numerical Study

In this section, we conduct simulation studies to illustrate the performance of the proposed method on pairwise link and hyperlink predictions on network. In particular, we investigate two scenarios

where either the incorporated hyperlinks are observed or inferred through the observed pairwise links.

3.5.1 Study 1: Link Predictions When the Hyperlinks are Observed

In the first simulation study, we consider the network generated from a true underlying latent space model and compare the performance of various methods under different missing rates for the observed pairwise links and hyperlinks.

Suppose there are $N = 50$ nodes in the network and the latent factors $\boldsymbol{Z} = \{Z_i\}_{i=1}^N$ for these nodes are generated from the following 5-dimensional multivariate Gaussian distribution such that $Z_i \sim \mathcal{N}(\boldsymbol{\mu}_k, 0.5\boldsymbol{I}),\ 10k+1 \leq i \leq 10(k+1)(k = 0, 1, 2, 3, 4)$, where

$$\begin{aligned}
\boldsymbol{\mu}_1 &= (0.5, 0.5, 0.5, 0.5, -0.5), \\
\boldsymbol{\mu}_2 &= (0.5, -0.5, 0.5, -0.5, 0.5), \\
\boldsymbol{\mu}_3 &= (-0.5, 0.5, -0.5, 0.5, -0.5), \\
\boldsymbol{\mu}_4 &= (-0.5, -0.5, -0.5, -0.5, 0.5), \\
\boldsymbol{\mu}_5 &= (0, 0, 0, 0, 0).
\end{aligned}$$

Denote the undirected and unweighted adjacent matrix for the generated network as Y. Given the latent factors $\boldsymbol{Z}$, a specific pairwise link connecting node i and node j is generated from the Bernoulli distribution based on the concordance of their latent factors such that

$$Y_{ij} \sim Bern(P_{ij}),\ P_{ij} = \exp(\langle Z_i, Z_j \rangle)/\{1 + \exp(\langle Z_i, Z_j \rangle)\}. \tag{3.12}$$

After generating random network based on latent factors $\boldsymbol{Z}$, we randomly choose some pairs of nodes in the network and mask their connecting status to serve as missing links with associated

missing rate of 20%. Therefore, the generated adjacency matrix is formulated as:

$$Y_{ij} = \begin{cases} 1, \text{link exists between i and j,} \\ 0, \text{no link between i and j ,} \\ -1, \text{observation is missing.} \end{cases} .$$

To simplify the formulation, we consider 3-order hyperlinks, where each hyperlink Y_{ijk} is directly generated from the same latent factors $\boldsymbol{Z}$ to mimic a case where hyperlinks are directly observed. That is,

$$Y_{ijk} \sim Bern(P_{ijk}), P_{ijk} = \sigma(Z_i Z_j^T + Z_i Z_k^T + Z_j Z_k^T + \sum_{r=1}^{5} Z_{ir} Z_{jr} Z_{kr}), \tag{3.13}$$

where $\sigma(\cdot)$ can be a logistic link. This follows the previous hierarchical modeling such that the hyperlinks are also consistent with the pairwise links. Notice the hyperlinks and pairwise links are conditional independent to each other given the shared latent factors $\boldsymbol{Z}$, and can serve as independent samples in generating probability P_{ij} and P_{ijk}.

The training dataset consists of both the pairwise links and hyperlinks which are generated directly. Specifically, we randomly split the set of observed pairwise links $\{Y_{ij}|Y_{ij} \neq -1, 1 \leq i \neq j \leq N\}$ into training, validation and test sets with the proportion at 50%, 15% and 15%, respectively. For generating the training and test datasets for hyperlinks, we follow a more strict sampling procedure instead of randomly sampling. The training hyperlinks are sampled from the testing pairwise links. That is, for each testing pairwise link Y_{ij}, we collect hyperlinks to form a candidate set $\Omega_{ij} = \{Y_{ijk}|Y_{jk} \neq -1, Y_{ik} \neq -1\}$ to ensure that the pairwise links within hyperlinks set are partially observed. Then we randomly select a hyperlink $Y_{ijk'}$ from this candidates set Ω_{ij} corresponding to testing link Y_{ij}. Finally, we choose these sampled hyperlinks as training dataset $\mathcal{Y} = \{Y_{ijk'}\}$ for the joint loss function (3.12).

Following this procedure, the size of the training 3-order hyperlinks is about 1000 given a network with node size at $N = 50$, which accounts for about 0.4% of all possible 3-order hyperlinks. Compared with the number of pairwise links, the number of hyperlinks is relatively sparser. This is consistent with the real application settings in the sense that the high-order or multi-way relation

is typically more difficult to observe and costs more to verify in contrast to the pairwise links. In terms of testing hyperlinks, we collect the 3-order hyperlinks such that pairwise links within hyperlinks are observed $\{Y_{ijk}|Y_{ij} \neq -1, Y_{ik} \neq -1, Y_{jk} \neq -1, 1 \leq i \neq j \neq k \leq N\}$. The rationale for such selection is that given two-way relations are observed, inferring the hyperlinks is more accurate and interpretable.

We provide a sample network generated from the above procedure on training and testing sets in Figure 3.3. The solid red lines and shaded circles represent the pairwise links and 3-order hyperlinks serving as training data respectively. The dashed lines and circles are the pairwise links and hyperlinks serving as the testing links. Notice that for each testing pairwise link, it is included in at least one training hyperlink. Similarly, for each testing hyperlink, the within pairwise links are observed. Modeling and utilizing the hierarchical dependency between pairwise links and hyperlinks is essential for the proposed method since it enables us to borrow information from both pairwise links and hyperlinks to improve the prediction.

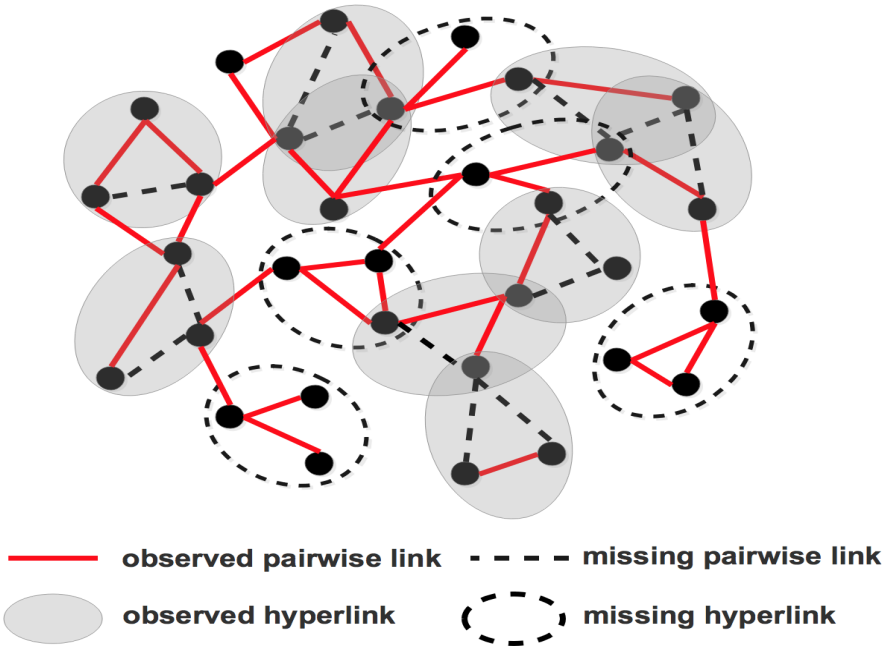


Figure 3.3: Sample network generated through the hierarchical relation between pairwise links and hyperlinks

To investigate the performance of incorporating hyperlinks on improving the link prediction, we compare six different methods. The first three methods are based on the proposed framework. Specifically, the first one obtains the estimation of the latent factors $\boldsymbol{Z}$ through the loss function

in (3.1), which is equivalent to utilize only observed pairwise links for embedding, and is denoted as **PLE**. The second method estimates the latent factors $\boldsymbol{Z}$ through the loss function defined in (3.2), which amounts to encoding only the training hyperlinks into $\boldsymbol{Z}$, and is denoted as **HLE**. The proposed method is denoted as **PLE+HLE** that incorporates both pairwise links and hyperlinks to jointly estimate $\boldsymbol{Z}$ through the proposed loss function (3.3). In addition, we also compare the proposed method with other three popular and state-of-art network embedding methods such as learning graph representations with global structural information (GraRep), large-scale information network embedding (LINE) and scalable feature learning for networks (Node2Vec). In general, they encode the observed pairwise or high-order relations into the node-wise latent factors through decomposing a series of graphical laplacian matrices with different orders, or encouraging $\boldsymbol{Z}$ to be conformed with the similarity among nodes obtained from biased random walk on network.

We obtain node-wise embedding estimation through different methods mentioned above and then predict the probability of testing links to be connected following (3.6) and (3.7). The performance of prediction is measured by the AUC (area under the ROC curve) indicator for the testing links.

Table 3.1: The AUC of link predictions on test data (test) and the entire network (global), **HLE** and **PLE** are based on parts of the proposed joint loss function.

	Link Prediction			
	Pairwise Link		**3rd-order Hyperlink**	
METHOD	test	global	test	global
PLE+HLE	**0.733**	**0.771**	**0.765**	**0.765**
HLE	0.651	0.643	0.647	0.649
PLE	0.666	0.747	0.661	0.649
GreRep	0.598	0.570	0.576	0.565
LINE	0.568	0.509	0.546	0.542
Node2Vec	0.486	0.497	0.501	0.499

The prediction results are illustrated in the Table 3.1, which shows that the proposed joint embedding method consistently outperforms other methods in terms of achieving higher AUC score. Specifically, although estimating latent factors through pairwise links or hyperlinks separately might not be adequately capture both two-way relations and three-way relations, the proposed

joint estimation approach is able to capture pairwise or hyperlinks simultaneously. For the pairwise link prediction, the proposed joint estimation (**PLE+HLE**) achieves about 10% improvement compared with pairwise link embedding (**PLE**) on test dataset. This indicates that the proposed method is capable of borrowing additional information for a two-way relation from the associated three-way hyperlinks, and obtain a more accurate pairwise similarity estimation. In addition, the joint embedding also achieves 18% improvement on hyperlink prediction compared to using hyperlink embedding (**HLE**) alone, indicating that the proposed method can make a better prediction for hyperlinks by utilizing the two-way relations within hyperlinks. This strategy of borrowing mutual information is built on utilizing the hierarchical dependency between pairwise links and hyperlinks through latent factors sharing between (3.1) and (3.2). Meanwhile, the joint embedding strategy takes advantage of this dependency to encode both two-way relations and multi-way relations into latent factors $\boldsymbol{Z}$, hence achieve better performance in both prediction tasks.

Compared with other three embedding methods, the proposed method achieves about 18% to 52% improvement for pairwise link prediction, and about 32% to 50% improvement for hyperlink prediction on test datasets. The improvement of the proposed method demonstrates that the high-order proximity captured by hyperlinks is crucial for inferring potential pairwise links. In terms of hyperlink prediction, the collection of two-way relations alone is not sufficient to determine the underlying multi-way relations. In addition, although all these methods indeed extract high-order relations through random walks or multi-step transition probabilities, the high-order information are essentially decomposed and formulated as pairwise relations, which leads to information loss for the high-order concordance and subgroups. The performance on pairwise link predictions from competing methods is generally inferior as they are not designed to incorporate the hierarchical dependency between pairwise links and hyperlinks. In addition, the simulated networks have relatively large randomness on the location of links, which do not possess specific structures represented in any of competing methods, as the competing methods are more oriented to specific application contexts.

3.5.2 Study 2: Link Prediction with the Hyperlinks Inferred

In numerical study 1, the incorporated hyperlinks are directly sampled from the correctly specified

generating process. However, in practice, hyperlinks are typically difficult to detect or verify as the complexity of high-order relations grows fast as the number of involved nodes increases. Consequently, the hyperlinks can be sparse, or almost unobserved, and are likely misspecified.

In this subsection, we generate the simulation setting to mimic the situation where partial hyperlinks are misspecified or hyperlinks are not directly observed but inferred from the observed pairwise links. The inferred hyperlinks could be misspecified due to the error propagation from misspecified pairwise links, randomness in sampling, and the discrepancy between the two-way relations and multi-way relations. Therefore, the prediction using inferred hyperlinks tends to be less power than that of incorporating observed hyperlinks. However, the intrinsic hierarchical dependency and latent factor sharing between hyperlinks and pairwise links still benefit the recovery if partial high-order information from an appropriate inference procedure since the proposed data augment procedure collects more information than the observed two-way relations.

We first investigate the performance of proposed methods which incorporate partially misspecified hyperlinks. The training dataset and testing dataset are generated following the same setting as in numerical study 1. The pairwise links are generated by (3.12) and randomly split into training set and testing set after removing the 20% of links, and treat them as missing. The third-order hyperlinks are generated through (3.4) and sampled following the previous one to form hierarchical relations with pairwise links. We also randomly sample 30% hyplerinks in the training dataset and flip their signs. That is

$$Y'_{ijk} = 0 \text{ if } Y_{ijk} = 1,$$
$$Y'_{ijk} = 1 \text{ if } Y_{ijk} = 0.$$

We replace these selected $\{Y_{ijk}\}$ by $\{Y'_{ijk}\}$ to mimic the scenario of misspecified hyperlinks. The prediction results are illustrated in Table 3.2, showing that the joint embedding of pairwise links and partially misspecified hyperlink still leads to better prediction performance. Specifically, under the 20% missing rate, the improvement of joint modeling pairwise links only (**PLE**) is 19% on the pairwise links and 28% on the hyperlinks for the test set. While the improvement over the only hyperlinks (**HLE**) is 7% on the pairwise links and 11% on the hyperlinks for the test set, which indicates the improvement is relatively robust even with misspecified high-order information. In

Table 3.2: The AUC on the test data and entire network **with 30% misspecified hyperlinks.**

		Missing 20%	Missing 50%
		Pairwise Links	
test dataset	**PLE+HLE**	0.756	0.691
	HLE	0.708	0.642
	PLE	0.635	0.593
entire network	**PLE+HLE**	0.833	0.770
	HLE	0.718	0.650
	PLE	0.818	0.739
		3rd-order hyperlinks	
test dataset	**PLE+HLE**	0.813	0.744
	HLE	0.738	0.660
	PLE	0.634	0.626
entire network	**PLE+HLE**	0.800	0.719
	HLE	0.732	0.650
	PLE	0.598	0.576

addition, when the number of observed pairwise links increases when the missing rate is 20%, and the number of hyperlinks from training set increases accordingly, both of all these methods improve. In particular, the improvement of **PLE** on hyperlink predictions demonstrates the benefits of introducing the dependency between pairwise link and hyperlink. In addition, the proposed joint embedding (**PLE+HLE**) and hyperlink embedding (**HLE**) gain a more significant improvement compared to the pairwise link embedding (**PLE**), which confirm that hyperlinks encode additional high-order relations that may not be substituted by two-way relations.

The following numerical experiment investigate performance of the proposed method when incorporating inferred hyperlinks instead of observed hyperlink. The hyperlinks in the training set are not sampled from the underlying true model (3.13). Instead, we generate training hyperlinks from observed pairwise links through the inferred procedure introduced in Section 3.4. The performance of the link predictions are illustrated in Table 3.3.

Table 3 shows that the proposed joint embedding method still achieves better performance compared with embedding methods using partial information. Specifically, the joint embedding achieve 8% improvement compared with the inferred hyperlink embedding (**HLE**) in both pairwise link and hyperlink predictions. In addition, two-way relations can only recover partial multi-way relations, and therefore the inferred hyperlinks suffer from the high-order information loss.

Table 3.3: The AUC of link predictions by incorporating inferred three-order hyperlinks on test data (test) and the entire network (global)

	Link Prediction			
	Pairwise Link		**3rd-order Hyperlink**	
	test	global	test	global
PLE+HLE	**0.690**	**0.830**	**0.715**	**0.702**
HLE	0.635	0.653	0.659	0.645
PLE	0.680	0.820	0.703	0.688

Compared with the pairwise link embedding (**PLE**), the joint embedding method achieves slightly improvement in pairwise links and hyperlinks even though we do not observe any hyperlink information. However, the improvement is limited due to that the dependency is only moderate, and restrictive information borrowed from dependency between pairwise links and hyperlinks.

3.6 Real Data Application

In this section, we apply the proposed joint link prediction method to the Facebook social circles network dataset (`https://snap.stanford.edu/data/egonets-Facebook.html`), which contains the ego-network. Each ego denotes a specific user in Facebook, and his or her associated ego-network is the social network corresponding to this user's friends in Facebook. One of the important attributes of ego-network is that it contains the social circles defined by the user, which leads to subgroups among people in the ego-network. Currently, social circle is a common functionality for many popular social medium such as Facebook, Google and Twitter. The purpose of introducing social circle is to allow users to organize their own social network to mitigate the 'information overload' through filtering contents or status updates posted by friends in specific groups. In addition, it allows users to protect their privacy by hiding or sharing personal information for specific groups of friends.

The major distinction between social circles and traditional social communities is that social circles are in general highly overlapped and can be hierarchically nested, therefore people in the ego-network generally might belong to multiple circles. In addition, unlike the social communities identified with the dense internal connections, social circles are formulated through specific

attributes of friends selected by the user. For example, a user might cluster his or her social networks according to categories such as college friends, high school friends, department friends or colleagues from the department.

Currently, social media adopts two methods to formulate social circles in the ego-network. The first one requires the user to manually group the people, which is time consuming and cannot be updated automatically when the user adds more friends. The second approach categorizes social circles through identifying people sharing common predefined attributes. However, it fails to incorporate the user's individual preferences and suffers from the missing profile information. In this subsection, we apply the proposed approach and investigate the performance of learning the user-specified clustering through network embedding.

The network includes 224 nodes as users and 6384 undirected pairwise links as friend relationships. In addition, there are 14 overlapped circles within the network, which are formulated according to the similarity of social features among people defined by the user, such as students of common universities, sports teams and relatives. These circles are in general with large size such that the on average each circle contains 40 nodes and the largest circle contains 201 nodes. The ego-network is illustrated in Figure 3.4.

We first randomly split the pairwise links in the ego-network into training, validation and test sets with the proportion of $50\%, 35\%$ and 15%. Incorporating of the multi-way information requires more elaborate preprocess. We first discard the most non-informative social circles, i.e., the largest social circle with size 201 such that almost all the people are within it that some specific attributes shared by almost all the people, which is not of our interest since we focus on the affect of multi-way relation on the subgroup structures, while the large social circle is non-informative in differentiating subgroup-specific attributes.

Next we encode the social circles as observed multi-way relations and incorporate it into the proposed method. In most of hyperlink embedding literature in adopting tensor representation, the nonhomogeneous hyperlink size is a non-trivial problem as a possible combination of nodes connected by a hyperlink can increases exponentially as the size of hyperlinks increases. One solution is to concatenate multiple tensors with different orders where each encoded hyperlink has a specific size. However, this strategy is infeasible in the context of ego-network as the total number of circles is small while their sizes are large. Consequently, this kind of representation strategy leads

to a set of ultra-sparse high-order tensors, which could suffer from high computational instability during the decomposition.

To solve this problem, we propose an alternative solution and apply it to encode the social circles. Instead of directly encoding the original social circles through high-order tensor, we first decompose the social circles into the three-way hyperlinks according to the following rule: $Y_{ijk} = 1$ if people $\{i, j, k\}$ are in the same original social circles, and $Y_{ijk} = 0$ otherwise, where Y_{ijk} is the third-order hyperlink connecting status among $\{i, j, k\}$. Therefore, the multi-way relations from the original social circles are represented as third-order hyperlinks, and then encoded into a third-order tensor. The local multi-way relations in the original social circles are captured by the three-way relations and the global subgroups can be approximately recovered by the collections of the overlapped third-order hyperlinks. After the decomposition, the original social circles are transformed into a set of third-order hyperlinks which might downweigh those important bridge links lying from the overlaps of multiple social circles. In practice, we select those overlapping three-order hyperlinks for training set. Through this preprocess, instead of using multiple high-order sparse tensors, we obtain a third-order dense tensor, to facilitate the downstream analysis. The numerical results also demonstrate that the proposed encoding strategy provides an adequate approximation for original high-order relations. Finally, we follow the procedure in the simulation study 1 to sample the training hyperlinks. The size of hyperlink training set is about 4,400 after the preprocess steps, and the proportions for positive instances ($\{(i, j, k)|Y_{ijk} = 1\}$) and negative instances ($\{(i, j, k)|Y_{ijk} = 0\}$) are balanced.

We also investigate the performance of the proposed method and other competing network embedding methods on predicting both two-way and multi-way relations on the ego-network. The two-way relation prediction is measured by the AUC on the testing pairwise links. Specifically, we predict the pairwise link Y_{ij} through $P(Y_{ij} = 1) = \frac{\exp(Z_i Z_j^T)}{1+\exp(Z_i Z_j^T)}$, where $\boldsymbol{Z} = \{Z_i\}_{i=1}^N$ are the estimated latent factors.

In addition to the pairwise link prediction, we also investigate the performance of hyperlink prediction to evaluate whether the social circle information has been encoded into nodes' latent factors. Instead of directly predicting the original social circles, we predict the joint memberships of specific m nodes, i.e., whether they belong to the same social circle or not, and compare the result with original circles. This task is still challenging in the sense that although the training

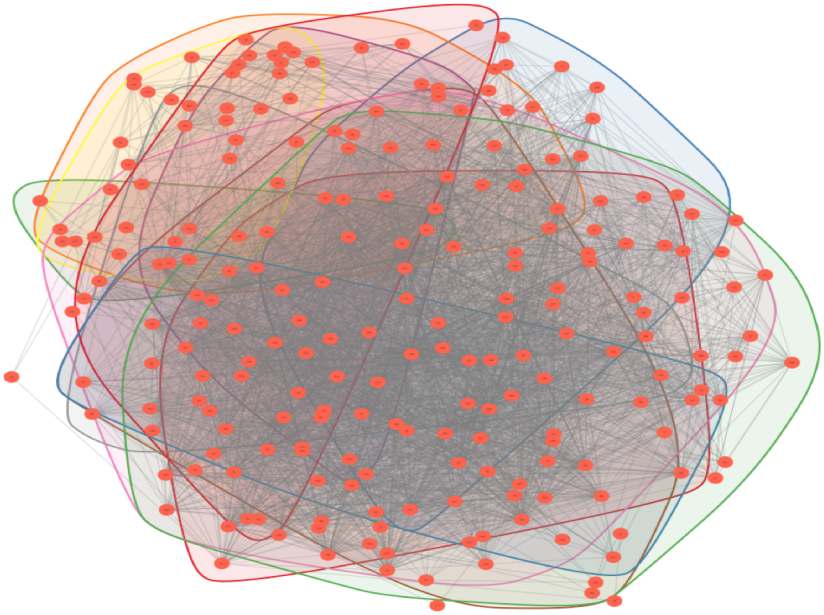


Figure 3.4: The ego network in the facebook dataset, where the social circles are marked as polygons with different colors

hyperlinks is third-order, the prediction on the hyperlinks for the test has high order beyond three. We choose the order of testing hyperlinks as $m = 6, 10$, and the prediction is based on estimated $\boldsymbol{Z}$ through

$$P(A_{i_1 i_2 \cdots i_m} = 1) = \frac{\exp(\sum_{(i,j)\in\{1,\cdots,m\}} Z_i Z_j^T)}{1 + \exp(\sum_{(i,j)\in\{1,\cdots,m\}} Z_i Z_j^T)}.$$

For the proposed method PLE+HLE, and the counter methods of using PLE and HLE separately, the rank of $\boldsymbol{Z}$ is chosen at $r = 5$. For the tuning parameters in other three methods, we adopt the strategies recommended by the original papers or the packages. In addition, our empirical study shows that their performance are not sensitive to the tuning parameters.

The comparison of performance is illustrated in Table 3.4. For the two-way relation prediction, the proposed joint embedding method has the best performance. It achieves about 6.5% improvement over the best existing method GreRep, 37% improvement over LINE and 70% improvement over Node2vec. Again, the real data analyses show the importance and benefits of borrowing high-order relation information for the two-way relation predictions. The incorporated third-order hyperlinks recover the underlying subgroup structure, and the estimated latent features characterize the subgroup attributes. If partial subgroup information are recovered, the two-way relations

Table 3.4: AUC of link prediction for ego-network

	Link Prediction			
	Pairwise Link		**6-order Hyperlink**	**10-order Hyperlink**
	test	global	test	test
PLE+HLE	0.850	0.868	0.750	0.920
HLE	0.792	0.788	0.740	0.890
PLE	0.825	0.868	0.599	0.569
GreRep	0.798	0.815	0.770	0.505
LINE	0.624	0.598	0.510	0.754
Node2Vec	0.498	0.496	0.492	0.485

from the subgroup level provide supplementary information regarding possibility of their friendship besides the concordance between their own individual latent features. In addition, the inferior performance of LINE and Node2Vec might result from their random-walk nature, which leads to biased and inefficient estimation of latent factors $\boldsymbol{Z}$ on the relatively sparse ego-network as the pairwise links only account for 13% on the total possible friendships.

In terms of the multi-way relation prediction, the proposed method is the second best for 6-order hyperlink prediction. However, it is closed to GreRep with 2.6% difference in AUC. For the prediction of the 10-order hyperlinks, the proposed method performs the best, and achieves 22% improvement over the best existing method LINE. It is noticeable that although the partial embedding method PLE and the GreRep perform well for the two-way relation prediction, they do not possess the consistent performance for hyperlink predictions, which indicates that the social circles indeed encode the significant high-order relations information which could not be represented through two-way relations. Therefore, it demonstrates the importance to incorporate the high-order information for predicting multi-way relation. In addition, although we only incorporate three-way relations into the proposed method, the estimated latent factors $\boldsymbol{Z}$ lead to a good prediction for the higher-order relations beyond the third order, implying that the proposed multi-way relation decomposition provides an adequate approximation for high-order relations, and the estimated $\boldsymbol{Z}$ are able to encode the subgroup information.

3.7 Discussion

In this chapter, we propose a new network link prediction method. The major innovation of the proposed method is to incorporate the multi-way relation into the network embedding process and therefore jointly embeds the hyperlinks and pairwise links. It allows the node-wise latent factors to encode both of the pairwise similarity to their neighbourhood and induce cohesive high-order subgroup. In addition, the proposed method formulates a hierarchical modeling for the link generating process to introduce the dependency between pairwise links and hyperlinks. In terms of estimating latent factors, the link dependency allows borrowing the mutual information between pairwise links and hyperlinks such that prediction for both pairwise links and hyperlinks can be improved. In term of model interpretability, the link dependency reflects the principle that high-order interaction among nodes in networks in general are generated from the low-order interactions.

In theory, we establish the consistency of the node-wise latent factors estimator based on the proposed joint embedding loss function as the number of observed links increasing. In addition, we show that the convergence rate can be improved through incorporating the hyperlinks. If the hyperlinks are directly observed as independent samples, then the improvement depends on the size of observed hyperlinks. On the other hand, the improvement from incorporating inferred hyperlinks depends on the size of observed pairwise links.

In this chapter, we only consider two scenarios that all the hyperlinks are either directly observed or inferred from pairwise links. However, in many real applications, the training data likely contains a small number of hyperlinks and relatively abundant pairwise links. It is worth of further exploration on developing an inference procedure which is capable of learning the complex relations between hyperlinks and pairwise links. It is possible to generate low-bias hyperlinks from a set of pairwise links, and improve the prediction performance through the proposed joint embedding procedure.

3.8 Notations and Proofs

3.8.1 Proof of Theorem 3.1

We first need to define the metric in the parameter space Θ in terms of the loss function. Denote

$$\tilde{l}_{joint}(\boldsymbol{\Theta}; \boldsymbol{Y}, \mathcal{Y}) = l_{pair}(\boldsymbol{\Theta}; \boldsymbol{Y}) + l_{hyper}(\boldsymbol{\Theta}; \mathcal{Y}),$$

and correspondingly $f(e; \boldsymbol{\Theta}) = \tilde{l}_{joint}(e; \boldsymbol{\Theta}) - \tilde{l}_{joint}(e; \boldsymbol{\Theta_0})$ where $e \in \boldsymbol{Y} \cup \mathcal{Y}$. Specially, considering the metric in the parameter space

$$\rho^2(\Theta_0, \Theta) = E(\tilde{l}_{joint}(\boldsymbol{\Theta_0}; \boldsymbol{Y}, \mathcal{Y}) - \tilde{l}_{joint}(\boldsymbol{\Theta}; \boldsymbol{Y}, \mathcal{Y})).$$

Given the binary edges, it is easy to testify that $f(e; \boldsymbol{\Theta})$ is bounded such that $|f(Y, \Theta)| \leq T$ for some $T > 0$. Denote the $\boldsymbol{Y}$ and $\mathcal{Y}$ as the set of observed pairwise links and hyperlinks generated from underlying model, and $\boldsymbol{E} = \boldsymbol{Y} \cup \mathcal{Y}$ as the set of observed links.

In this paper we choose the L_2 loss for estimating both pairwise generating probability θ_{ij} and three-order hyperlink generating probability θ_{ijk} with $l(\boldsymbol{\Theta}) = (Y - \boldsymbol{\Theta})^2$, where $\boldsymbol{\Theta} = \{\theta_{ij}, \theta_{ijk}\}$. Define the distance on the parameter space $\mathcal{S}$ as $\rho(\boldsymbol{\Theta}, \boldsymbol{\Theta_0}) = K^{1/2}(\boldsymbol{\Theta}, \boldsymbol{\Theta_0})$ where $\boldsymbol{\Theta_0}$ is the true parameter set,

$$\begin{aligned}
K(\boldsymbol{\Theta}, \boldsymbol{\Theta_0}) &= \frac{1}{N^2 + N^3}(\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} E(l(Y_{ijk}, \theta_{ijk}) - l(Y_{ijk}, \theta_{0,ijk})) \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{N} E(l(Y_{ij}, \theta_{ij})^2 - l(Y_{ij}, \theta_{0,ij})), \\
V(\boldsymbol{\Theta}, \boldsymbol{\Theta_0}) &= \frac{1}{N^2 + N^3}(\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} Var(l(Y_{ijk}, \theta_{ijk}) - l(Y_{ijk}, \theta_{0,ijk})) \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{N} Var(l(Y_{ij}, \theta_{ij})^2 - l(Y_{ij}, \theta_{0,ij})),
\end{aligned}$$

where Y_{ij} and Y_{ijk} represents the random links generating under the true model. Since the $\boldsymbol{\Theta_0}$ is the true parameters then $K(\boldsymbol{\Theta}, \boldsymbol{\Theta_0}) \geq 0$ and $K(\boldsymbol{\Theta}, \boldsymbol{\Theta_0}) = 0$ only if $\boldsymbol{\Theta} = \boldsymbol{\Theta_0}$. Then define the

distance on the parameter space as $\rho(\mathbf{\Theta_0}, \mathbf{\Theta}) = K^{1/2}(\mathbf{\Theta}, \mathbf{\Theta_0})$. Through calculation we have

$$K(\mathbf{\Theta}, \mathbf{\Theta_0}) = \frac{\|\mathbf{\Theta} - \mathbf{\Theta_0}\|^2}{N^3 + N^2},$$

$$Var(\mathbf{\Theta}, \mathbf{\Theta_0}) = \frac{4Var(Y)\|\mathbf{\Theta} - \mathbf{\Theta_0}\|^2}{N^3 + N^2} \leq \frac{\|\mathbf{\Theta} - \mathbf{\Theta_0}\|^2}{N^3 + N^2}.$$

We then split the parameter space $\mathcal{S}$ through: $A(k_1, k_2) = \{\mathbf{\Theta} \in \mathcal{S} : k_1 \leq \rho(\mathbf{\Theta_0}, \mathbf{\Theta}) \leq 2k_1, J(\mathbf{\Theta}) \leq k_2\}$, and $F(k_1, k_2) = \{\tilde{l}_{joiint}(\mathbf{\Theta}) - \tilde{l}_{joint}(\mathbf{\Theta_0}) : \mathbf{\Theta} \in A(k_1, k_2)\}$. The main procedure of the proof is to satisfies that the proposed loss function satisfying the three assumptions for the corollary 2 in (Shen, 1994) and estimate the corresponding metric entropy to determine the best convergence rate for the minimizer of the proposed loss function.

We first verify the proposed loss function satisfying the assumption 1 and assumption 2 for corollary 2 in Shen (1994) [105]. By definition, we have

$$\sup_{A(k_1,k_2)} V(\mathbf{\Theta_0}, \mathbf{\Theta}) \leq c_8 k_1^2 = c_8 k_1^2 \left\{1 + \left(k_1^2 + k_2\right)^{\beta_1}\right\}$$

with $\beta_1 = 0$. For either the pairwise or hyperlinks Y, we have with some constant c, $|\theta_0 - \theta|^2 \leq c \operatorname{Var}\{l(\theta, Y) - l(\theta_0, Y)\}$. Furthermore,

$$|l(\theta, Y) - l(\theta_0, Y)| = |\theta_0 - \theta| \cdot |2Y - \theta_0 - \theta|.$$

Define a new random variable $W = |2Y - \theta - \theta_0|$. Notice both Y and θ are bounded in $[0, 1]$. Therefore $\sup |W|$ is bounded.

Then we verify that for a constant $c > 0$, we have $\sup_{A(k_1,k_2)} \|\mathbf{\Theta_0} - \mathbf{\Theta}\|_{\sup} \leq c\left(k_1^2 + k_2\right)^{\beta_2}$ with $\beta_2 \in [0, 1)$. Recall $f(e; \mathbf{\Theta(Z)}) := \tilde{l}_{joint}(e; \mathbf{\Theta}) - \tilde{l}_{joint}(e; \mathbf{\Theta_0})$ is also function of the latent factors $\boldsymbol{Z}$ and the total number of parameters is $\gamma = rN$. Given the assumption (C1) and the fact that $f(e; Z)$ is a smooth function of Z, we have $f(e; Z) \in W_p^{\infty}[-C, C]^{\gamma}$ where W_p^{∞} is a Sobolev space with $p \geq 2$. From the definition of $\rho(\mathbf{\Theta_0}, \mathbf{\Theta})$ and $A(k_1, k_2)$, we have $\|f(e; Z)\|_2 = \rho(\mathbf{\Theta_0}, \mathbf{\Theta})$ which is bounded. Based on the lemma 2 in (Shen, 1994), it follows that $\|f(e; Z)\|_{\infty} = \|\mathbf{\Theta_0} - \mathbf{\Theta}\|_{\infty}$ is bounded. Then $\sup_{A(k_1,k_2)} \|\Theta_0 - \Theta\|_{\sup} \leq c\left(k_1^2 + k_2\right)^{\beta}$ with $\beta_2 = 0$.

To introduce the bracket metric entropy, let $\mathcal{N}(\varepsilon, n) = \left\{f_1^l, f_1^u, \ldots, f_n^l, f_n^u\right\}$ be a set of functions

from the L_2 space such that $\max_{1\leq i\leq n}\left\|fi^u - f_i^l\right\|_2 \leq \varepsilon$. Suppose for any loss function in sliced parameter space $\mathcal{F}(k_1, k_2)$, there exists a set of $\{f_i^l, f_i^u, i = 1, \cdots, n\}$ such that almost surely

$$f_i^l \leq \tilde{l}_{joint}(\boldsymbol{\Theta}) - \tilde{l}_{joint}(\boldsymbol{\Theta_0}) \leq f_i^u.$$

Then the L_2 metric entropy with bracketing f is defined as $H(\varepsilon, \mathcal{F}(k_1, k_2)) = \log\{n : \min \mathcal{N}(\varepsilon, n)\}$.

Finally we estimate the smallest ϵ satisfying the assumption 3 through estimating the metric entropy $H(\varepsilon, \mathcal{F})$ first. Let the parameter $\omega = \frac{\alpha}{\gamma} = \infty$ where $\alpha = \infty$ is the smoothness of f, hence $p\omega = \infty > 1$. Based on the Theorem 5.2 in [23], with a constant c the metric entropy satisfying

$$H\left(\varepsilon_{|\boldsymbol{E}|}, \mathcal{F}_2(k_1, k_2)\right) \leq c\varepsilon_{|\boldsymbol{E}|}^{-1/\omega} = c.$$

Then for fixed k_1 and k_2, we estimate the order of left hand on assumption 3. Based on $\beta_1 = 0, \beta_2 = 0$ we have

$$\psi(k_1, k_2) = \sqrt{c}\frac{U - L}{L} \leq \frac{\varepsilon_{|\boldsymbol{E}|} - \lambda_{|\boldsymbol{E}|}}{\lambda_{|\boldsymbol{E}|}}.$$

To achieve the best rate with smallest ϵ, the metric entropy needs to satisfy $\psi \sim |\boldsymbol{E}|^{1/2}$. Accordingly, the best rate for ϵ is $\varepsilon_{|\boldsymbol{E}|} \sim \frac{1}{|\boldsymbol{E}|^{1/2}}$ with $\varepsilon_{|\boldsymbol{E}|} \sim \lambda_{|\boldsymbol{E}|}^{1/2}$. Given the observed pairwise links and observed hyperlinks are independent samples from the underlying generating probability Θ_0. The result in Theorem 3.4.1 then follows by applying Corollary 2 of Shen (1994) [105].

3.8.2 Proof of Theorem 3.2

Denote the $\boldsymbol{Y}$ as the set of observed pairwise links generated from underlying model, $\mathcal{Y}$ as the set of inferred hyperlinks which are unbiased to the true generating probability, and $\boldsymbol{E} = \boldsymbol{Y} \cup \mathcal{Y}$ as the augmented set of links. Given the dependency between pairwise links and inferred hyperlinks, the classical Bernstein's inequality is not applicable and therefore need the following lemma.

lemma 3.1. *For any threshold $t > 0$ and previously defined function $f(e; \boldsymbol{\Theta})$, we have*

$$\mathbb{P}\left[\frac{1}{|\boldsymbol{E}|}\sum_{e\in\boldsymbol{E}}(f(e;\boldsymbol{\Theta}) - \mathbb{E}\left[f(e;\boldsymbol{\Theta})\right]) > t\right] \leq \exp\left(-C\frac{|\boldsymbol{E}|t^2}{T^2\max_i(d_i)}\right),$$

where $|\boldsymbol{E}|$ is the number of links, d_i is the degree of node i and C is a positive constant.

Proof: consider the number of pairwise links associated with an end point node i as m_i, then $\sum_{i=1}^{N} m_i = |\boldsymbol{Y}|$. The total number of inferred hyperlinks are $\sum_{i=1}^{N} m_i^2 \geq \frac{|\boldsymbol{Y}|^2/N - |\boldsymbol{Y}|}{2}$, where N is the number of nodes in the network. Therefore, the total number of links $|\boldsymbol{E}| \geq \sum_{i=1}^{N} m_i^2 + |\boldsymbol{Y}| = \frac{|\boldsymbol{Y}|^2/N + |\boldsymbol{Y}|}{2}$. We introduce the concept of smallest proper cover introduced in Christoph, et. al (2018) [63] and it can be testified that the size of proper covers for $\boldsymbol{E}$ is $\boldsymbol{O}(\max_i(d_i))$. Then the result follows from the Theorem 2 in Christoph, et. al (2018) [63].

We first introduce the following large deviation inequality. Note that the following result is different to other similar convergence property for estimator in term of empirical process as most of them consider the case such as Wong and Shen (1994) [117] where the samples are independent observed, while the sample links are correlated to each other in our setting.

We first introduce the metric entropy of a class $\mathcal{F}$. Given $\epsilon > 0, p > 0$, denote

$$\mathcal{N}(\epsilon, \mathcal{F}) := \min \left\{ k : \text{ there exist } f_1, \cdots, f_K \in \mathcal{F} \text{ such that } \min_{i \leq k} \|f - f_i\|_p < \varepsilon \text{ for all } f \in \mathcal{F} \right\},$$

and $H_p(\epsilon, \mathcal{F}) := \log \mathcal{N}(\epsilon, \mathcal{F})$. In the following we specify $\mathcal{F} = \{f(e, \boldsymbol{\Theta}), \boldsymbol{\Theta} \in \mathcal{S}\}$. We have the following lemma:

lemma 3.2. *Consider the function $f(e; \boldsymbol{\Theta})$ in Lemma 1, $a \in (0, 1)$ and $M > 0$. Define t_0 through $H_\infty(t_0, \mathcal{F}) = \frac{|\boldsymbol{E}| M^2}{4T^2 \max_i(d_i)}$ and $s = \frac{aM}{64}$. If*

$$\int_s^{t_0} H_\infty^{1/2}(u, \mathcal{F}) du \leq |\boldsymbol{E}|^{1/2} M a^{3/2} / 2^8,$$

then

$$\mathbb{P}^* \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{|\boldsymbol{E}|} \sum_{e \in \boldsymbol{E}} (f(e; \boldsymbol{\Theta}) - \mathbb{E}\left[f(e; \boldsymbol{\Theta}) \right]) \right) > M \right] \leq 3 \exp\left(-C \frac{(1-a)|\boldsymbol{E}| M^2}{T^2 \max_i(d_i)} \right),$$

where C is a positive constant and $\mathbb{P}^*$ is outer measure.

Proof: the proof for lemma 2 follows a similar chain argument for Theorem 2.1 in Alexander (1984) [7] with the $\psi(M, n)$ replaced by $\frac{|\boldsymbol{E}| M^2}{T^2 \max_i(d_i)}$.

Next we define a distance $\rho(,)$ on the parameter space $\mathcal{S}$ such that $\rho(\theta,\theta_0)=K^{1/2}(\theta,\theta_0)$ where $K(\theta,\theta_0)$ is introduced in Theorem 4.1. Notice the argument in this proof can be extended to other loss function satisfying certain regularization conditions instead of $f(e;\mathbf{\Theta})$. We continue the proof with the L_2 loss function adopted in this paper.

We divide the parameter space $\mathcal{S}$ into pieces as $A\left(k_1,k_2\right)=\{\theta\in\mathcal{S}:k_1\leq\rho\left(\theta_0,\theta\right)\leq 2k_1, J(\theta)\leq k_2\}$ and $\mathcal{F}_1\left(k_1,k_2\right)=\left\{\tilde{l}_{joint}(e;\boldsymbol{\theta})-\tilde{l}_{joint}(e;\boldsymbol{\theta_0}):\theta\in A\left(k_1,k_2\right)\right\}$ where $k_1>0, k_2>0$. We assume that the function space $\mathcal{F}$ satisfying the following Assumption 1 and later we show that the proposed loss function satisfies Assumption 1. In the following proof $c_i's$ and $d_i's$ denote positive constants.

Assumption 1: the metric entropy satisfying that $\sup_{k_1\geq 1,k_2\geq 1}\psi_2\left(k_1,k_2\right)\leq c_1|\boldsymbol{E}|^{1/2}$ where

$$\psi_2\left(k_1,k_2\right)=\int_L^U H_\infty^{1/2}\left(u,\mathcal{F}_1\left(k_1,k_2\right)\right)du/L,$$

where $U=d_1\varepsilon\left(k_1^2+k_2\right)^{1/2}$ and $L=d_2\lambda\left(k_1^2+k_2\right)$. The Assumption 1 intuitively controls the size of function space $\mathcal{F}_1(k_1,k_2)$.

Denote $\tilde{l}(\theta,e)=\tilde{l}_{joint}(e;\theta)-\lambda J(\theta)$. Let

$$\begin{aligned}\nu_n\left(\tilde{l}(\theta,e)-\tilde{l}\left(\theta_0,e\right)\right)&=|\boldsymbol{E}|^{-1}\sum_{e\in\boldsymbol{E}}\left(\tilde{l}\left(\theta,e\right)-\tilde{l}\left(\theta_0,e\right)-E\left(\tilde{l}\left(\theta,e\right)-\tilde{l}\left(\theta_0,e\right)\right)\right)\\&=\nu_n\left(\tilde{l}_{joint}(\theta,e)-\tilde{l}_{joint}\left(\theta_0,e\right)\right).\end{aligned}$$

For $i=1,2,\ldots,j=0,1,\ldots$, denote

$$\begin{aligned}A_{i,j}=\{\theta\in\Theta:2^{i-1}\varepsilon\leq\rho\left(\theta_0,\theta\right)<2^i\varepsilon, 2^{j-1}\max\left(J\left(\theta_0\right),1\right)\leq J(\theta)\\<2^j\max\left(J\left(\theta_0\right),1\right)\}.\end{aligned}$$

Without loss of generality, we assume $\max(\lambda|\boldsymbol{E}|,\epsilon)\leq 1$. Therefore, given $\max(J(\theta_0),1)\lambda\leq c\epsilon^2$

for any $i, j \geq 1$ we have

$$\begin{aligned}\inf_{A_{i,j}}\left[K\left(\theta_0,\theta\right)+\lambda_n\left(J(\theta)-J\left(\theta_0\right)\right)\right] &\geq \left(2^{i-1}\varepsilon\right)^2+\lambda_n\left(2^{j-1}-1\right)J\left(\theta_0\right)\\ &\geq c_2\lambda[(2^{i-1})^2+(2^{j-1}-1)J(\theta_0)],\end{aligned}$$

and

$$\inf_{A_{i,0}}\left[K\left(\theta_0,\theta\right)+\lambda_n\left(J(\theta)-J\left(\theta_0\right)\right)\right] \geq \left(2^{i-1}\varepsilon\right)^2-\lambda_n J\left(\theta_0\right) \geq c_3\left(2^{i-1}\varepsilon\right)^2.$$

Denote $M(i,j)=c_2\lambda\left[\left(2^{i-1}\right)^2+\left(2^{j-1}-1\right)J\left(\theta_0\right)\right]$ and $M(i)=c_3\left(2^{i-1}\varepsilon\right)^2$ then we have

$$\begin{aligned}I =&P^*\left(\sup_{\{\rho(\theta_0,\theta)\geq\varepsilon,\theta\in\mathcal{S}\}}|\boldsymbol{E}|^{-1}\sum_{e\in\boldsymbol{E}}\left(\tilde{l}\left(\theta,e\right)-\tilde{l}\left(\theta_0,e\right)\right)\geq-\varepsilon^2/2\right)\\ =&\sum_{i,j=1}^{\infty}P^*\left(\sup_{A_{i,j}}\nu_n\left(\tilde{l}_{joint}(\theta,e)-\tilde{l}_{joint}\left(\theta_0,e\right)\right)\geq M(i,j)\right)\\ &+\sum_{i=1}^{\infty}P^*\left(\sup_{A_{i,0}}\nu_n\left(\tilde{l}_{joint}(\theta,e)-\tilde{l}_{joint}\left(\theta_0,e\right)\right)\geq M(i)\right)\\ =&I_1+I_2.\end{aligned}$$

We first bound I_1 with utilizing the previous Lemma 2 on each parameter space slice $A_{i,j}$. Notice that $M(i,j)$ is in the order $c_4\lambda(2^{2i}+2^j)$. Denote t_1 as $H_\infty(t_1,A_{i,j})=|\boldsymbol{E}|M^2(i,j)$. Based on Assumption 1, we have $H_\infty(U_{i,j},A_{i,j})\leq c_5|\boldsymbol{E}|M^2(i,j)$ where $U_{i,j}=d_1\epsilon(2^{i-1}\epsilon+2^j)$ which leads to $t_1\leq U_{i,j}$. Also denote $s_1=c_6M_{i,j}$. Therefore, by Assumption 1 on the metric entropy we have for any $i\geq 1, j\geq 1$

$$\int_{s_1}^{t_1}H_\infty^{1/2}\left(u,\mathcal{F}_2\left(2^i\varepsilon,2^j\right)\right)du/M(i,j)\leq\int_L^{U_{i,j}}H^{1/2}\left(u,\mathcal{F}_2\left(2^i\varepsilon,2^j\right)\right)du/L\leq d_3|\boldsymbol{E}|^{1/2}.$$

Therefore in each sliced parameter space $A_{i,j}$, we are able to use Lemma 2 to have

$$\begin{aligned} I_1 &\leq 3\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\exp\left(-c_7|\boldsymbol{E}|M(i,j)^2/\max_i(d_i)\right) \\ &\leq 3\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\exp\left(-c_8|\boldsymbol{E}|\lambda^2\left[\left(2^{i-1}\right)^2+2^{j-1}\right]^2/\max_i(d_i)\right) \\ &\leq 3\exp\left(-c_9|\boldsymbol{E}|\lambda^2/\max_i(d_i)\right)/\left[1-\exp\left(-c_9|\boldsymbol{E}|\lambda^2/\max_i(d_i)\right)\right]. \end{aligned}$$

Following the similar discussion we can obtain the same bound for I_2, then

$$I \leq 6\exp\left(-c_{10}|\boldsymbol{E}|\lambda^2/\max_i(d_i)\right)/\left[1-\exp\left(-c_{10}|\boldsymbol{E}|\lambda^2/\max_i(d_i)\right)\right] \leq 7\exp\left(-\frac{c_{10}|\boldsymbol{E}|\lambda^2}{\max_i(d_i)}\right).$$

It follows that

$$P^*\left(\sup_{\{\rho(\theta_0,\theta)\geq\varepsilon,\theta\in\mathcal{S}\}}|\boldsymbol{E}|^{-1}\sum_{e\in\boldsymbol{E}}\left(\tilde{l}\left(\theta,e\right)-\tilde{l}\left(\theta_0,e\right)\right)\geq-\varepsilon^2/2\right)\leq 7\exp\left(-c_{10}|\boldsymbol{E}|\lambda^2/\max_i(d_i)\right),$$

where $\boldsymbol{E}$ is the set consist of both observed pairwise links and underlying unbiased hyperlinks. Recall that $\boldsymbol{E}=\Omega_{\boldsymbol{Y}}\cup\Omega_{\mathcal{Y}}$ where $\Omega_{\boldsymbol{Y}}$ is the set of observed pairwise links and $\Omega_{\mathcal{Y}}$ is the set of the inferred hyperlinks such that its size depends on the size of $\Omega_{\boldsymbol{Y}}$ to reflect the facts that the more pairwise links are observed the more hyperlinks we are able to infer. Given the assumption (C2) on the inference error for hyperlinks:

$$\frac{1}{|\Omega_{\mathcal{Y}}|}\sum_{Y_{ijk}\in\Omega_{\mathcal{Y}}}|\hat{Y}_{ijk}-Y_{ijk}|\leq O(\epsilon^2),$$

where $\hat{Y}_{ijk}$ is the inferred hyperlink from an inference procedure based on the Y_{ij},Y_{ik},Y_{jk}. The interpretation of the assumption is that the more pairwise links are observed the closer between the distribution of hyperlinks from true model and that of inferred hyperlinks. Based on the L_2 loss function adopted in this paper and the fact that Y_{ijk}'s corresponding generating probability θ_{ijk} are

bounded, we have:

$$\frac{1}{|\Omega_{\mathcal{Y}}|}\sum_{Y_{ijk}\in\Omega_{\mathcal{Y}}}|\tilde{l}(\theta,\hat{Y}_{ijk})-\tilde{l}(\theta,Y_{ijk})|\leq O(\epsilon^2).$$

Then we replace the unbiased hyperlinks Y_{ijk} by the inferred version $\hat{Y}_{ijk}$ therefore to incorporate the inference error. Notice that the estimation $\hat{\theta}$ satisfies

$$\begin{aligned}|\boldsymbol{E}|^{-1}\sum_{Y_{ij},\hat{Y}_{ijk}\in\boldsymbol{E}}\tilde{l}\left(\hat{\theta},Y_{ij},\hat{Y}_{ijk}\right)&\geq\sup_{\theta\in\Theta}|\boldsymbol{E}|^{-1}\sum_{Y_{ij},\hat{Y}_{ijk}\in\boldsymbol{E}}\tilde{l}\left(\theta,Y_{ij},\hat{Y}_{ijk}\right)-a\\&\geq|\boldsymbol{E}|^{-1}\sum_{Y_{ij},\hat{Y}_{ijk}\in\boldsymbol{E}}\tilde{l}\left(\theta_0,Y_{ij},\hat{Y}_{ijk}\right)-a.\end{aligned}$$

Therefore, given $a\leq O(\epsilon^2)$ we have

$$\begin{aligned}&P(\rho\left(\hat{\theta},\theta_0\right)\geq\varepsilon)\leq P^*(\sup_{\{\rho(\theta_0,\theta)\geq\varepsilon,\theta\in\mathcal{S}\}}|\boldsymbol{E}|^{-1}\sum_{Y_{ij},\hat{Y}_{ijk}\in\boldsymbol{E}}\left(\tilde{l}\left(\theta,Y_{ij},\hat{Y}_{ijk}\right)-\tilde{l}\left(\theta_0,Y_{ij},\hat{Y}_{ijk}\right)\right)\geq-a)\\&\leq P^*(\sup_{\{\rho(\theta_0,\theta)\geq\varepsilon,\theta\in\mathcal{S}\}}|\boldsymbol{E}|^{-1}\sum_{Y_{ij},Y_{ijk}\in\boldsymbol{E}}\left(\tilde{l}\left(\theta,Y_{ij},Y_{ijk}\right)-\tilde{l}\left(\theta_0,Y_{ij},Y_{ijk}\right)\right)\geq\\&\qquad-a-\frac{c_1 1}{|\Omega_2|}\sum_{Y_{ijk}\in\Omega_2}|\tilde{l}(\theta,\hat{Y}_{ijk})-\tilde{l}(\theta,Y_{ijk})|)\\&\leq P^*(\sup_{\{\rho(\theta_0,\theta)\geq\varepsilon,\theta\in\mathcal{S}\}}|\boldsymbol{E}|^{-1}\sum_{Y_{ij},Y_{ijk}\in\boldsymbol{E}}\left(\tilde{l}\left(\theta,Y_{ij},Y_{ijk}\right)-\tilde{l}\left(\theta_0,Y_{ij},Y_{ijk}\right)\right)\geq-c\epsilon^2))\\&\leq 7\exp\left(-c_{10}|\boldsymbol{E}|\lambda^2/\max_i(d_i)\right).\end{aligned}$$

If $\max(J(\theta_0),1)\lambda\leq c\epsilon^2$, then

$$P(\rho\left(\hat{\theta},\theta_0\right)\geq\varepsilon)\leq 7\exp\left(-c_{12}|\boldsymbol{E}|\epsilon^4/\max_i(d_i)\right),$$

when $\lambda\sim\epsilon^2$. Otherwise, if $\max(J(\theta_0),1)\lambda\geq c\epsilon^2$, then

$$P(\rho\left(\hat{\theta},\theta_0\right)\geq\lambda^{1/2})\leq P(\rho\left(\hat{\theta},\theta_0\right)\geq\varepsilon)\leq 7\exp\left(-c_{10}|\boldsymbol{E}|\lambda^2/\max_i(d_i)\right).$$

Combining both two situations, we have

$$P(\rho\left(\hat{\theta},\theta_0\right)\geq\eta)\leq 7\exp\left(-c_{13}|\boldsymbol{E}|\eta^4/\max_i(d_i)\right), \tag{3.14}$$

where $\eta=\max\{\lambda^{1/2},\epsilon\}$ with ϵ being the smallest value satisfying Assumption 1. In the following we testify that the proposed joint loss function satisfying the Assumption 1 and the smallest ϵ can be estimated by the metric entropy inequality. For the L_2 loss function we use the metric $K(\boldsymbol{\Theta},\boldsymbol{\Theta_0})$ on the parameter space $\mathcal{S}$ defined in Theorem 4.1 such that $\rho(\boldsymbol{\Theta_0},\boldsymbol{\Theta})=K^{1/2}(\boldsymbol{\Theta},\boldsymbol{\Theta_0})$. Recall $f(e;\boldsymbol{\Theta(Z)}):=\tilde{l}_{joint}(e;\boldsymbol{\Theta})-\tilde{l}_{joint}(e;\boldsymbol{\Theta_0})$ is also function of the latent factors $\boldsymbol{Z}$ and the total number of parameters is $\gamma=rN$. Given the assumption (C1) and the fact that $f(e;Z)$ is a smooth function of Z, we have $f(e;Z)\in W_p^{\infty}\left[-C,C\right]^{\gamma}$ where W_p^{∞} is a Sobolev space with $p\geq 2$.

Finally we estimate the smallest ϵ satisfying the Assumption 1 through estimating the metric entropy $H_{\infty}\left(\varepsilon,\mathcal{F}\right)$. Let the parameter $\omega=\frac{\alpha}{\gamma}=\infty$ where $\alpha=\infty$ is the smoothness of $f(e;Z)$, hence $p\omega=\infty>1$. Based on the Theorem 5.2 in [23], with a constant c the metric entropy for $\mathcal{F}_1(k_1,k_2)$ satisfying

$$H_{\infty}\left(\varepsilon,\mathcal{F}_1(k_1,k_2)\right)\leq c_{14}\varepsilon^{-1/\omega}=c_{14}\varepsilon^{-0}=c_{14}.$$

Then for fixed k_1 and k_2, the order of left hand on Assumption 1,

$$\psi\left(k_1,k_2\right)=\sqrt{c}\frac{U-L}{L}\leq\frac{\varepsilon-\lambda}{\lambda}.$$

Based on the constraint that $\psi\left(k_1,k_2\right)\leq c_1|\boldsymbol{E}|^{1/2}$, the metric entropy needs to satisfy $\psi\sim|\boldsymbol{E}|^{1/2}$ to achieve the smallest ϵ. Accordingly, the best rate for ϵ is $\varepsilon_{|\Omega|}\sim\frac{1}{|\boldsymbol{E}|^{1/2}}$ with $\varepsilon_{|\boldsymbol{E}|}\sim\lambda_{|\boldsymbol{E}|}^{1/2}$. Recall from Lemma 1 the fact that $|\boldsymbol{E}|\geq\frac{|\boldsymbol{Y}|^2/N+|\boldsymbol{Y}|}{2}$, the result in Theorem 3.4.2 then follows from (3.14).

Bibliography

[1] Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.

[2] Agarwal, S., Branson, K., and Belongie, S. (2006). Higher order learning with graphs. In *Proceedings of the 23rd international conference on Machine learning* 17–24. ACM.

[3] Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V., and Smola, A. J. (2013). Distributed large-scale natural graph factorization. In *Proceedings of the 22nd international conference on World Wide Web* 37–48. ACM.

[4] Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014.

[5] Al Hasan, M., Chaoji, V., Salem, S., and Zaki, M. (2006). Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*.

[6] Al Hasan, M. and Zaki, M. J. (2011). A survey of link prediction in social networks. In *Social network data analytics* 243–275. Springer.

[7] Alexander, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *The Annals of Probability* 1041–1067.

[8] Alexander-Bloch, A., Lambiotte, R., Roberts, B., Giedd, J., Gogtay, N., and Bullmore, E. (2012). The discovery of population differences in network community structure: new methods and applications to brain functional networks in schizophrenia. *Neuroimage*, 59(4):3889–3900.

[9] Amelio, A., Mangioni, G., and Tagarelli, A. (2019). Modularity in multilayer networks using redundancy-based resolution and projection-based inter-layer coupling. *IEEE Transactions on Network Science and Engineering*.

[10] Amini, A. A., Chen, A., Bickel, P. J., Levina, E., et al. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122.

[11] Anagnostopoulos, A., Kumar, R., and Mahdian, M. (2008). Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data mining* 7–15. ACM.

[12] Arias-Castro, E., Chen, G., Lerman, G., et al. (2011). Spectral clustering based on local linear approximations. *Electronic Journal of Statistics*, 5:1537–1587.

[13] Bader, G. D. and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1):2.

[14] Bahadur, R. R. (1959). A representation of the joint distribution of responses to n dichotomous items. Technical report.

[15] Balakrishnan, S., Wainwright, M. J., Yu, B., et al. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120.

[16] Ball, B., Karrer, B., and Newman, M. E. (2011). Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3).

[17] Barigozzi, M., Fagiolo, G., and Garlaschelli, D. (2010). Multinetwork of international trade: A commodity-specific analysis. *Physical Review E*, 81(4):046104.

[18] Bechtel, J. J., Kelley, W. A., Coons, T. A., Klein, M. G., Slagel, D. D., and Petty, T. L. (2005). Lung cancer detection in patients with airflow obstruction identified in a primary care outpatient practice. *Chest*, 127(4):1140–1145.

[19] Belkin, M. and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems* 585–591.

[20] Betzel, R. F., Bertolero, M. A., Gordon, E. M., Gratton, C., Dosenbach, N. U., and Bassett, D. S. (2018). The community structure of functional brain networks exhibits scale-specific patterns of variability across individuals and time. *bioRxiv* 413278.

[21] Bhattacharyya, S. and Chatterjee, S. (2018). Spectral clustering for multiple sparse networks: I. *arXiv preprint arXiv:1805.10594*.

[22] Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*.

[23] Birman, M. S. and Solomyak, M. Z. (1967). Piecewise-polynomial approximations of functions of the classes w_pˆα. *Matematicheskii Sbornik*, 115(3):331–355.

[24] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10).

[25] Bordenave, C., Lelarge, M., and Massoulié, L. (2015). Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on* 1347–1357. IEEE.

[26] Bretto, A. (2013). Hypergraph theory. *An introduction. Mathematical Engineering. Cham: Springer*.

[27] Bui-Xuan, B.-M. and Jones, N. S. (2014). How modular structure can simplify tasks on networks: parameterizing graph optimization by fast local community detection. *Proc. R. Soc. A*, 470(2170).

[28] Cao, S., Lu, W., and Xu, Q. (2015). Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM international on conference on information and knowledge management* 891–900. ACM.

[29] Celisse, A., Daudin, J.-J., Pierre, L., et al. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899.

[30] Chen, X. and Yang, Y. (2018). Hanson-wright inequality in hilbert spaces with application to k-means clustering for non-euclidean data. *arXiv preprint arXiv:1810.11180*.

[31] Cheng, J., Levina, E., Wang, P., and Zhu, J. (2014). A sparse ising model with covariates. *Biometrics*, 70(4):943–953.

[32] Chitra, U. and Raphael, B. J. (2019). Random walks on hypergraphs with edge-dependent vertex weights. *arXiv preprint arXiv:1905.08287*.

[33] Choi, D. S., Wolfe, P. J., and Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284.

[34] Corsini, P. and Leoreanu, V. (2013). *Applications of hyperstructure theory*, volume 5. Springer Science & Business Media.

[35] Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., et al. (2013). The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477.

[36] Das, J. and Yu, H. (2012). Hint: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology*, 6(1):92.

[37] Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183.

[38] De Domenico, M., Nicosia, V., Arenas, A., and Latora, V. (2015). Structural reducibility of multilayer networks. *Nature communications*, 6.

[39] Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics* 49–93.

[40] Donetti, L. and Munoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(10).

[41] Estrada, E. and Rodríguez-Velázquez, J. A. (2006). Subgraph centrality and clustering in complex hyper-networks. *Physica A: Statistical Mechanics and its Applications*, 364:581–594.

[42] Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. (2015). The reactome pathway knowledgebase. *Nucleic acids research*, 44(D1):D481–D487.

[43] Fortunato, S. and Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41.

[44] Garcia, J. O., Ashourvan, A., Muldoon, S., Vettel, J. M., and Bassett, D. S. (2018). Applications of community detection techniques to brain graphs: Algorithmic considerations and implications for neural function. *Proceedings of the IEEE*, 106(5):846–867.

[45] Geng, J., Bhattacharya, A., and Pati, D. (2018). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, (just-accepted):1–32.

[46] Ghoshal, G., Zlatić, V., Caldarelli, G., and Newman, M. E. (2009). Random hypergraphs and their applications. *Physical Review E*, 79(6):066118.

[47] Handcock, M. S. (2003). Statistical models for social networks. *Dynamic social network modeling and analysis. National Academies Press, Washington*.

[48] Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354.

[49] Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems* 657–664.

[50] Hoff, P. D. (2018). Additive and multiplicative effects network models. *arXiv preprint arXiv:1807.08038*.

[51] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.

[52] Hristova, D., Noulas, A., Brown, C., Musolesi, M., and Mascolo, C. (2016). A multilayer approach to multiplexity and link prediction in online geo-social networks. *EPJ Data Science*, 5(1):24.

[53] Jaakkola, T. (2001). Tutorial on variational approximation methods. *Advanced Mean Field Methods: Theory and Practice*.

[54] Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing* 665–674. ACM.

[55] Jalili, M., Orouskhani, Y., Asgari, M., Alipourfard, N., and Perc, M. (2017). Link prediction in multiplex online social networks. *Royal Society open science*, 4(2):160863.

[56] Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1).

[57] Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998.

[58] Kim, N., Wilburne, D., Petrović, S., and Rinaldo, A. (2016). On the geometry and extremal properties of the edge-degeneracy model. *arXiv preprint arXiv:1602.00180*.

[59] King, A. D., Pržulj, N., and Jurisica, I. (2004). Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020.

[60] Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of complex networks*, 2(3):203–271.

[61] Klamt, S., Haus, U.-U., and Theis, F. (2009). Hypergraphs and cellular networks. *PLoS computational biology*, 5(5):e1000385.

[62] Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L., and Zhang, P. (2013). Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940.

[63] Lampert, C. H., Ralaivola, L., and Zimin, A. (2018). Dependency-dependent bounds for sums of dependent random variables. *arXiv preprint arXiv:1811.01404*.

[64] Latouche, P., Birmele, E., and Ambroise, C. (2012). Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1):93–115.

[65] Lauritzen, S., Rinaldo, A., and Sadeghi, K. (2018). Random networks, graphical models and exchangeability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):481–508.

[66] Lawrence, E., Michailidis, G., Nair, V. N., and Xi, B. (2006). Network tomography: A review and recent developments. In *Frontiers in statistics* 345–366. World Scientific.

[67] Le, C. M., Levin, K., Levina, E., et al. (2018). Estimating a network from multiple noisy realizations. *Electronic Journal of Statistics*, 12(2):4697–4740.

[68] Le, C. M., Levina, E., Vershynin, R., et al. (2016). Optimization via low-rank approximation for community detection in networks. *The Annals of Statistics*, 44(1):373–400.

[69] Lee, K.-M., Yang, J.-S., Kim, G., Lee, J., Goh, K.-I., and Kim, I.-m. (2011). Impact of the topology of global macroeconomic network on the spreading of economic crises. *PloS one*, 6(3).

[70] Levin, K., Lodhia, A., and Levina, E. (2019). Recovering low-rank structure from multiple networks with unknown edge distributions. *arXiv preprint arXiv:1906.07265*.

[71] Li, D., Xu, Z., Li, S., and Sun, X. (2013). Link prediction in social networks based on hypergraph. In *Proceedings of the 22nd International Conference on World Wide Web* 41–42. ACM.

[72] Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

[73] Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170.

[74] Mariadassou, M., Robin, S., Vacher, C., et al. (2010). Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics*, 4(2):715–742.

[75] Massoulié, L. (2014). Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing* 694–703. ACM.

[76] Menon, A. K. and Elkan, C. (2011). Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases* 437–452. Springer.

[77] Meunier, D., Lambiotte, R., and Bullmore, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in neuroscience*, 4:200.

[78] Montanari, A. and Saberi, A. (2010). The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201.

[79] Mossel, E., Neeman, J., and Sly, A. (2015). Consistency thresholds for the planted bisection model. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing* 69–75. ACM.

[80] Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*, 9(5):471.

[81] Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

[82] Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3).

[83] Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2).

[84] Newman, M. E. and Reinert, G. (2016). Estimating the number of communities in a network. *Physical Review Letters*, 117(7).

[85] Nobile, A. and Fearnside, A. T. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, 17(2):147–162.

[86] Ogburn, E. L., VanderWeele, T. J., et al. (2017). Vaccines, contagion, and social networks. *The Annals of Applied Statistics*, 11(2):919–948.

[87] Ossiander, M. (1987). A central limit theorem under metric entropy with l2 bracketing. *The Annals of Probability* 897–919.

[88] Ou, M., Cui, P., Pei, J., Zhang, Z., and Zhu, W. (2016). Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* 1105–1114. ACM.

[89] Pattison, P. and Robins, G. (2002). 9. neighborhood-based models for social networks. *Sociological Methodology*, 32(1):301–337.

[90] Paul, S. and Chen, Y. (2018). A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging. *arXiv preprint arXiv:1805.02292*.

[91] Pavlovic, D. M. (2015). *Generalised Stochastic Blockmodels and their Applications in the Analysis of Brain Networks*. PhD thesis, University of Warwick.

[92] Pearson, K. J. and Zhang, T. (2014). On spectral hypergraph theory of the adjacency tensor. *Graphs and Combinatorics*, 30(5):1233–1248.

[93] Pinheiro, C. A. R. (2012). Community detection to identify fraud events in telecommunications networks. *SAS SUGI Proceedings: Customer Intelligence*.

[94] Pržulj, N., Wigle, D. A., and Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340–348.

[95] Pu, L. and Faltings, B. (2012). Hypergraph learning with hyperedge expansion. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 410–425. Springer.

[96] Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., and Tang, J. (2018). Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* 459–467. ACM.

[97] Rahman, M. S., Dey, L. R., Haider, S., Uddin, M. A., and Islam, M. (2017). Link prediction by correlation on social network. In *Computer and Information Technology (ICCIT), 2017 20th International Conference of* 1–6. IEEE.

[98] Ramadan, E., Tarafdar, A., and Pothen, A. (2004). A hypergraph model for the yeast protein complex network. In *18th International Parallel and Distributed Processing Symposium, 2004. Proceedings.* 189. IEEE.

[99] Razick, S., Magklaras, G., and Donaldson, I. M. (2008). irefindex: a consolidated protein interaction database with provenance. *BMC bioinformatics*, 9(1):405.

[100] Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007a). An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191.

[101] Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007b). Recent developments in exponential random graph (p*) models for social networks. *Social networks*, 29(2):192–215.

[102] Rohe, K., Chatterjee, S., Yu, B., et al. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915.

[103] Saade, A., Krzakala, F., and Zdeborová, L. (2014). Spectral clustering of graphs with the bethe hessian. In *Advances in Neural Information Processing Systems* 406–414.

[104] Saldana, D. F., Yu, Y., and Feng, Y. (2017). How many communities are there? *Journal of Computational and Graphical Statistics*, 26(1):171–181.

[105] Shen, X. (1998). On the method of penalization. *Statistica Sinica* 337–357.

[106] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

[107] Song, H. H., Cho, T. W., Dave, V., Zhang, Y., and Qiu, L. (2009). Scalable proximity estimation and link prediction in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement* 322–335. ACM.

[108] Stanley, N., Shai, S., Taylor, D., and Mucha, P. J. (2016). Clustering network layers with the strata multilayer stochastic block model. *IEEE Transactions on Network Science and Engineering*, 3(2):95–105.

[109] Tang, W., Lu, Z., and Dhillon, I. S. (2009). Clustering with multiple graphs. In *2009 Ninth IEEE International Conference on Data Mining* 1016–1021. IEEE.

[110] Torreggiani, S., Mangioni, G., Puma, M. J., and Fagiolo, G. (2018). Identifying the community structure of the food-trade international multi-network. *Environmental Research Letters*, 13(5):054026.

[111] Valverde-Rebaza, J. C. and de Andrade Lopes, A. (2012). Link prediction in complex networks based on cluster information. In *Advances in Artificial Intelligence-SBIA 2012* 92–101. Springer.

[112] Von Der Malsburg, C. (1994). The correlation theory of brain function. In *Models of Neural Networks* 95–119. Springer.

[113] Wang, S., Arroyo, J., Vogelstein, J. T., and Priebe, C. E. (2019). Joint embedding of graphs. *IEEE transactions on pattern analysis and machine intelligence*.

[114] Warnick, R., Guindani, M., Erhardt, E., Allen, E., Calhoun, V., and Vannucci, M. (2018). A bayesian approach for estimating dynamic functional network connectivity in fmri data. *Journal of the American Statistical Association*, 113(521):134–151.

[115] Waskiewicz, T. (2012). Friend of a friend influence in terrorist social networks. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[116] Wasserman, S., Faust, K., et al. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

[117] Wong, W. H., Shen, X., et al. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, 23(2):339–362.

[118] Wu, C. J. et al. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, 11(1):95–103.

[119] Zanin, M., Cano, P., Buldú, J. M., and Celma, O. (2008). Complex networks in recommendation systems. In *Proc. Second World Scientific and Eng. Academy and Soc. Int'l Conf. Computer Eng. and Applications* 120–124. Citeseer.

[120] Zhang, A. Y. and Zhou, H. H. (2017). Theoretical and computational guarantees of mean field variational inference for community detection. *arXiv preprint arXiv:1710.11268*.

[121] Zhang, J., Sun, W. W., and Li, L. (2018). Mixed-effect time-varying network model and application in brain connectivity analysis. *arXiv preprint arXiv:1806.03829*.

[122] Zhao, Y., Levina, E., Zhu, J., et al. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292.

[123] Zhou, D., Huang, J., and Schölkopf, B. (2007). Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in neural information processing systems* 1601–1608.

[124] Zhu, Y., Guan, Z., Tan, S., Liu, H., Cai, D., and He, X. (2016). Heterogeneous hypergraph embedding for document recommendation. *Neurocomputing*, 216:150–162.

[125] Zlatić, V., Ghoshal, G., and Caldarelli, G. (2009). Hypergraph topological quantities for tagged social networks. *Physical Review E*, 80(3):036118.